Epistemic Engineering

Uncovering the Logic of Deceivability and Meta-Induction

Christian J. Feldbacher-Escamilla

Version: 2019

Contents

for Mariana and Hagen

Introduction

In this chapter a short overview of topics, the tradition, and the structure of the book is provided. The book is about meta-induction *and how to engineer applications thereof in epistemology. It is in the wider tradition of formal learning theory and machine learning. And it consists of three main parts: a part introducing basic notions and results of the theory of meta-induction; a part on applying these results in the classical epistemic realm; and a part on applications in the realm of social epistemology.*

Epistemologists are concerned with the analysis of epistemic notions like *justification*, *truth*, *belief*, and *knowledge*. In doing so, they often put forward constraints and desiderata and try to design explications in agreement with these. Sometimes the desiderata turn out to be too demanding and one is able to show that no explication adequately suits them. In such a case one starts to fiddle around with the desiderata until an explication is no longer excluded for logical reasons. Then the designing procedure starts again. If one does so by help of tools and methods of the formal sciences, then this intuitive behaviour is alike that of an engineer who is faced with a task or problem and needs to devise a solution for it. In practice, it also quite often turns out that a task is too demanding in order to be reasonably accomplished. Then ways of conservatively modifying the task are investigated and devising a solution for the modified task starts again.

This book tries to tackle epistemological problems in the way just outlined; it deals with explicating the notion of *epistemic justification* broadly construed, and it does so by employing concepts and results of the theory of meta-induction. It takes up the sceptic's challenge that this notion is incoherent and argues for a conservative modification of the desiderata. The main idea is to explicate *justification* not as an *absolute*, but a *relative* notion: An inference method is not only justified, if it maximises epistemic values in absolute terms, but also, if it optimises them in comparison or relative to its alternatives. Such an instrumental approach is characteristic for rationality and rationalisation in the practical domain, in particular in engineering; and next to the choice of our formal tools, it is a reason why we call this kind of investigation an approach of *epistemic engineering*.

Before we outline the topics of the book in more detail, we want to locate our investigation very generally within the epistemic enterprise. As we have mentioned already above, epistemology is concerned with the key notions of *justification* (J), *truth* (T), *belief* (B), and *knowledge* (K), and we are focusing on *justification*. Whereas in the 20th century investigations of T, B, and K underwent dramatic developments, it seems that investigations of J were in a resigning slumber for a long time, and that only quite recently new developments in philosophy lead to an awakening within this domain also. Let us illustrate this by help of comparison. Willard van Orman Quine aptly claimed that "more than once in history the discovery of paradox has been the occasion for major reconstruction at the foundations of thought" (1966, p.3)—and without exaggerating it seems that also in epistemology one can indicate research foci quite well by reference to paradox or antinomy. Clearly, the list of epistemic paradoxes and antinomies is long (for an overview see Sorensen 2018; and more generally on paradoxes in philosophy see Clark 2002). However, it seems reasonable to attribute a great bulk of 20th century epistemic research to the following four: Regarding K, a case in point is *Gettier's paradox*: (KA1) One knows p if one has a justified true belief of p. (KA2) Justification is closed under known entailment. (KA3) There are *Gettier cases* where one lacks knowledge, although one has an (indirectly) justified true belief knowingly entailed by a (directly) justified belief which is false. Regarding B, it is *the lottery and preface paradox* which gained lots of attention: (BA1) It is not rational to believe a contradiction. (BA2) Rational belief is closed under conjunction. (BA3) Lockean bridging: One rationally believes p iff one's degree of belief in p passes a threshold. (BA4) There are cases where one's degree of belief in p and that in q passes such a threshold, but one's degree of belief in $p\&q$ does not. Regarding T, the *liar antinomy* is most prominent: (TA1) Sentence (TA1) is not true. And, finally, regarding J it is *Agrippa's trilemma*: (JA1) non-scepticism: Some beliefs are justified. (JA2) non-dogmatism: A belief is justified only by help of another justified belief. (JA3) non-coherentism: There are no circles in justification. (JA4) finitism: All justificatory chains are finite.

The first three paradoxes or antinomies have led to a "major reconstruction at the foundations of thought". Just for illustrative purpose, without claiming that this point of view is exhaustive, regarding K, *knowledge*, Gettier's paradox (KA1)–(KA3) triggered a whole industry of epistemic research which, amongst others, resulted in a division of the epistemic community in *internalists* and *externalists*; with respect to B, *belief*, the lottery- and preface paradox (BA1)–(BA4) led to far-ranging revisions of the qualitative notion of *belief*, some argued even for its abandonment, while others fundamentally revised the bridging principles between qualitative and quantitative notions of belief; regarding T and the liar (TA1) Alfred Tarski's theory of *truth* was epoch-making and, e.g., the distinction between object and meta language became indispensable for any primer in logic. But how about the fourth antinomy, the sceptic riddle of J, *justification* (JA1)–(JA4)?

Here it seems that many philosophers have thrown in the sponge and sceptics remained sceptic, dogmatists dogmatic, and coherentists simply coherent. Especially starting with the second third of the 20th century and the upcoming of theories of confirmation it seems that the *justification problem* was replaced by a *characterisation problem*, namely the problem of how to formally describe inductive inferences which one already accepted on intuitive grounds (see Vickers 2010).

An exception in this respect is Hans Reichenbach who was the first to think of justification in *relative* terms (see 1938, and 1940). Since focusing on achieving the best result *relative* to one's alternatives is a quite *pragmatic* turn in comparison of focusing on achieving the best result *simpliciter* or *absolutely*, this approach is also called a *pragmatic vindication*. Reichenbach's approach was pioneering, but failed to dramatically change our foundation of thought, because it seemed to prove *too much* and *too little* at the same time. Here is why: Reichenbach assumed that justified predictions are possible only for event series with a limit; however, there are infinitely many different competing methods which allow one to predict the limit of an event series correctly. I.e., in comparison with each other, all these methods are on a par and hence would be justified. So, Reichenbach's approach needs to be restricted in order to determine these cases better. On the other hand, his relative justification proves too little, because for obvious reasons one can be predictively successful also regarding event series without a limit, and so his approach needs to be extended in order to account also for these cases.

Now, both issues are addressed by two contemporary approaches, namely *formal learning theory* and the theory of *meta-induction* (see Henderson 2018, sect.7). Formal learning theory addresses the problem that a Reichenbachean vindication licenses too many inferences or sources of justification. The idea is that intuitively acceptable inferences allow for achieving further epistemic values which others do not. Further epistemic values might be *efficiency*, *quickness*, *minimal retractability*, etc. The formal learning approach traces back to Rudolf Carnap's programme of *logical probabilities* (1950/1962), was criticised early on from Reichenbach's student Hilary Putnam via computability considerations (1963), and was prominently taken up by Ray J. Solomonoff (1964); likewise as Putnam, but independently of him, E. Mark Gold had also found out about a connection between induction and computability while modelling language acquisition of a child (learner) whose mental structure encodes syntactic competence (learning target, hypothesis). Due to this influential model this approach gained its label *formal learning theory*. Contemporary perhaps most prominent is the formal learning approach represented by Kevin T. Kelly (1996). A survey of formal learning theory is provided in Osherson, Stob, and Weinstein (1986, subsequent to Gold), Sterkenburg (2018, subsequent to Putnam), and K. T. Kelly, Schulte, and Juhl (1997, subsequent to both).

The other route to the problem of epistemic justification is that of *meta-induction* which was sketched programmatically by Reichenbach himself, but carried out only quite recently by Schurz (2004, 2008, 2009, 2012a, 2012b, 2019), Schurz and Thorn (2016), Thorn and Schurz (2012, 2016, forthcoming), Arnold (2010), Feldbacher-Escamilla (2012, 2017), Feldbacher-Escamilla and Schurz (manuscript), Henderson (sect.7.3 2018), and Sterkenburg (2019). Here the idea is that one can get rid of Reichenbach's assumption of a limit in the series to be predicted simply by applying induction not, as is ordinarily done, on the object level, but on the meta level in form of extrapolating the success rates of prediction methods. There is an analytic result of the theory of meta-induction and a branch of *machine learning* which shows that such an extrapolation is guaranteed to be predictively successful in the long run *relative* to all available prediction methods; hence, meta-induction accounts for *relative* justification.

It is this route to epistemic justification that we take up in this book. We proceed in three steps: First, we present analytical results of meta-induction; second, we apply them to the core problem of epistemic justification, namely to the problem of justifying inductive practice; and third, we apply them to further problems of epistemic justification in the wider sense, particularly to the problem of justifying social epistemic practices. For this reason the book consists of three parts:

Part I Here we outline the programme of *epistemic engineering*. We show that, in contrast to traditional approaches, this programme aims at overcoming the problem of (*absolute*) epistemic justification via stressing the *relative* notion of justification and by this shifting the epistemic task from proving *an ideal* towards proving *an optimum* (chapter 1). In order to spell out this approach, we first describe the setting in detail (chapter 2); then, since scepticism—see (JA1) in Agrippa's trilemma—is the main challenger to epistemic justification and deceiving is its strongest weapon, we study the *logic of deceivability* and show a *blind spot* in the best deceiver's strategy which allows for proving the main optimality result to be applied in the rest of the book (chapter 3). Finally, we show that and how, given further constraints, this main result can be generalised (chapter 4).

Part II After having introduced all necessary engineering tools and the main optimality result, we present the meta-inductive solution to the core problem of epistemic justification, namely to the problem of justifying induction (chapter 5); once the core problem has been addressed, a *new riddle* seems to show up, namely the problem of how to justify induction without, at the same time, also justifying anti-induction; however, as we will argue, the assumptions underlying this follow-up problem are self-defeating and hence the follow-up problem does not undermine the solution of the core problem (chapter 6). In this part we also outline some bearing of this solution to the problem of justifying the two other forms of inferences used in science, namely abduction (chapter 7), and deduction (chapter 8).

Part III Finally, we are showing how justification in terms of optimisation allows also for new approaches to the problem of justification in the social epistemic realm. We go through the core topics of social epistemology and show how meta-inductive optimisation allows for a new approach to the problem of testimony (chapter 9), peer disagreement (chapter 10), judgement aggregation (chapter 11), and the wisdom of the crowds (chapter 12).

Before we delve into these issues, a brief note for the reader on what to expect and what to not expect is in place: The book is about applying results of the theory of meta-induction to philosophical problems; above we stated that there are two modern successors of Reichenbach's vindication approach to epistemic justification, namely *formal learning theory* and *meta-induction*. It is important to note that we consider here the latter only and *not* the former. In formal learning theory one approaches the problem of justifying induction the same way as a computer scientist conceives of computational problems which is to find an algorithm that is supposed to be guaranteed to output a correct answer on every possible input. The main problem formal learning theorists are after is the question which conditions have to be imposed on a series of events in order to guarantee for an algorithm, a computable object-inductive method, to achieve reliable results. In contrast to this, we almost never speak about *computability* and *complexity*—readers interested in this approach are referred to (K. T. Kelly 1996; Sterkenburg 2018). Furthermore, the focus of the first part of the book is on presenting *the main optimality result of the theory of meta-induction based on theorems about regret-based online learning of machine learning*; we aim at providing a didactically accessible presentation intended especially for a philosophical audience where basically high school mathematics (and a little bit of sweating) should suffice for verifying the result. We consider it to be *the* main result, because one can find (a reference to) it in primers and course manuals of so-called *online* machine learning. We do not provide much technical variations or new results in this respect—for such we refer the interested reader to (Schurz 2019). For didactic reasons we also aimed at presenting *the logic of deceivability* and its application as far as possible in combinatorial terms. Besides the didactics of the first part, the reader can and should particularly expect applications of the theory of meta-induction in this book. The application within the classical realm of epistemology (part II) is thought to provide self-contained support of and additions to the argumentation of Schurz (2019). The applications within the social epistemic realm (part III), on the other hand, provide a completely new expansion; we hope that they demonstrate further how far-reaching and fruitful the meta-inductive uncovering and *blindspotting* of the logic of deceivability is, and how by this the problem of epistemic justification can be overcome.

"I have the simplest tastes.
I am always satisfied with the best."
Oscar Wilde

Part I

Epistemic Engineering

Chapter 1

A Tour Through Contemporary Epistemology

This chapter provides a logical analysis of the problem of epistemic justification. Then merits and problems of the three classical approaches to this problem, namely foundationalism, coherentism, and infinitism are briefly discussed. Finally, the programme of naturalised epistemology, the programme whose stance on epistemic engineering this investigation uses as a take off, is sketched.

In this book we are mainly concerned with the problem of epistemic justification. The problem of epistemic justification is this: We think that we are justified in believing some propositions. However, such justification presupposes reasons, and such reasons, in turn, are in need of justification. So, given we are justified in believing something, we seem to face the following problem: Either we *dogmatically* stop at some point of the reasoning chain and do not provide any further reason. Or we *circularly* provide reasons. Or we never stop and go on with an *infinite regress*. This *trilemma of justification* is typically ascribed to Agrippa the Sceptic (first century CE) who mentioned two further *modes* or problems of justification: *dissent* (we will discuss this problem a little bit in chapter 10 on epistemic disagreement; this mode might be subsumed under *epistemic relativism* in the sense that if there is *reasonable disagreement*, adherents of the dissent camp might argue for a relativist position), and *relation* (which in the modern setting might be interpreted as some kind of epistemic contextualism, according to which the question of whether p is true or not is considered relative to a context). *Agrippa's trilemma* became quite influential in his aftermath during so-called *Pyrrhonism*, where many ancient academics turned to scepticism—they refused reasoning by dogmatism, circularity, and infinite regress, and—roughly speaking—bit the bullet of denying the existence of justified belief (see Sextus Empiricus 1999, book I, sect.1, p.45). We can explicate the trilemma even as an antinomy, which is—once one makes the non-sceptic stance explicit—a quadrilemma:

The Problem of Epistemic Justification.

(EJ1) Some beliefs are justified.
Schematically: $\exists x Jx$

(EJ2) No belief is justified unless some other belief serves as a justified reason for it.
Schematically: $\forall x(Jx \rightarrow \exists y(Jy \,\&\, yRx))$

(EJ3) If a belief is in the reason-ancestry for justifying another one, then the latter cannot serve (directly or indirectly) as a reason for justifying the former.
Schematically: $\forall xyz(xRy \rightarrow (\neg yRx \,\&\, yRz \rightarrow xRz))$
(I.e.: R is transitive and asymmetrical.)

(EJ4) The chain of justification via reasons is finite.
Schematically: The axioms on J and R from above have a finite model.

These four, prima facie plausible, principles of justification and reason are inconsistent: It is easy to see that the axioms in (EJ1)–(EJ3) have no finite model. Assume, e.g., a domain with three elements p_1, p_2, p_3 and let Jp_1 (in order to satisfy (EJ1)). Then, by (EJ2), we need to assume Jp_2 and p_2Rp_1 (w.l.o.g.—analogously for p_3; p_1Rp_1 is excluded by asymmetry of R according to (EJ3)). Now, since Jp_2, we need to assume further that Jp_3 and p_3Rp_2 (by (EJ2); p_1Rp_2 is, again, excluded by asymmetry of R). Now, since Jp_3, we need to assume that for some x: xRp_3 (by (EJ2)). $x \neq p_3$ and $x \neq p_2$ due to the asymmetry of R. Furthermore, $x \neq p_1$ due to asymmetry and transitivity of R according to (EJ3) (p_3Rp_2 and p_2Rp_1 implies p_3Rp_1 which excludes p_1Rp_3). Hence, in order to satisfy the axioms we need to assume a further element p_4. Induction on the number of elements of the domain completes the proof that no finite model satisfies (EJ1)–(EJ3).

In order to resolve the problem of justification, the following classical stances arose (see Van Cleve 2014, p.256):

- Scepticism: There is no justified belief, i.e. vs. (EJ1).
- Foundationalism (or dogmatism): There are some justified beliefs which are not in need of any justified reasons, i.e. vs. (EJ2).
- Coherentism: There are some beliefs serving (directly or indirectly) mutually as justifying reasons, i.e. vs. (EJ3).
- Infinitism: The chain of reasoning need not be finite, i.e. vs. (EJ4).

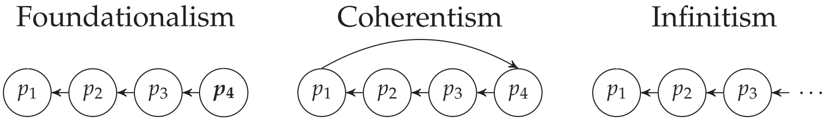

Figure 1.1: Traditional approaches to the problem of epistemic justification (the arrows represent the relation of providing a reason): Foundationalism assumes that there are reasons which are not in need of any further justification. Coherentism assumes that at least some reasons can provide mutual justification. Infinitism assumes that the chain of reasoning can be infinitely long. (For simplicity, transitivity is omitted in the graphical representation.)

The differences between the three main non-sceptical positions is illustrated with help of the schemata in figure 1.1.

Our formulation of the problem of epistemic justification is a little bit simplified. First of all, it seems that there are further possibilities to overcome this problem. Well known is, e.g., Karl R. Popper's formulation of the problem as *Fries' trilemma*:

> "The problem of the basis of experience has troubled few thinkers so deeply as Fries. He taught that, if the statements of science are not to be accepted *dogmatically*, we must be able to justify them. If we demand justification by reasoned argument, in the logical sense, then we are committed to the view that statements can be justified only by statements. The demand that all statements are to be logically justified [...] is therefore bound to lead to an *infinite regress*. Now, if we wish to avoid the danger of dogmatism as well as an infinite regress, then it seems as if we could only have recourse to *psychologism*, i.e. the doctrine that statements can be justified not only by statements but also by perceptual experience. Faced with this *trilemma*—dogmatism vs. infinite regress vs. psychologism—Fries, and with him almost all epistemologists who wished to account for our empirical knowledge, opted for psychologism. In sense-experience, he taught, we have 'immediate knowledge'." (Popper 2002b, p.75)

Popper mentions the *dogmatist* and the *infinite regress* approach to epistemic justification, however, he also notes another approach, namely *psychologism*, according to which the relation of justification is not only between statements, but also between perceptual experiences and statements. Regarding statements $p_1, p_2, \ldots$ the axioms in (EJ1)–(EJ4) still hold true. However, they do not hold for perceptual experiences $e_1, e_2, \ldots$. In the same line as foundationalism, psychologism needs to revise at least principle (EJ2). This might be as follows:

2′. No belief is justified unless some other belief serves as a justified reason for it or it is granted by experience.
Schematically: $\forall x(Jx \rightarrow (Ex \vee \exists y(Jy \,\&\, yRx)))$

It is easy to see that this is not only in the line of foundationalism, but an approach which was and still is wholeheartedly embraced by logical empiricists.

Different from this is Popper's programme of *falsificationism* which approaches epistemic justification in a way not covered by the argument above. The idea of falsificationism is that one need not to argue for a belief to be justified, but that justification is provisorily granted for any (in principle falsifiable) belief per default, i.e. from the start on. The task then is to rule out those beliefs that are reasonably considered to be unjustified, namely falsified. This principle of falsificationism is not in need of justification, since—according to falsificationism—it is justified per default. Since we do not discuss falsificationism any further in this investigation (with one exception: we will pick up falsificationism very briefly again when discussing the problem of induction in section 5.2), one of our simplifications of the problem of epistemic justification consists in bypassing *per default justification*.

Another simplification of our framing is that the problem of epistemic justification usually does not concern just simple chains of reasoning, but complete systems of belief. However, in order to make the difference between the main approaches to this problem clear, the simple chain structure suffices. Also, as Atkinson and Peijnenburg (2009, p.183) already mentioned, "the chain is a good starting point: it can help us to understand more realistic cases, which have been represented as trees, rafts, pyramids, teepees, houses of cards, cobwebs, or crossword puzzles, all of which have single chains as their elements."

Finally, and perhaps most importantly, we need to mention that this framing is only about the problem of justification on the lower level of basic propositions ($p_1, p_2, \ldots$). However, it is clear that justification is also about the rules which allow us to proceed from one proposition to another (represented as arrows in figure 1.1). Typically, these rules are considered to be either deductive or inductive, sometimes also abductive. The so-called *problem of higher level justification* concerns the question of how to justify these inference rules (see Schurz 2019, sect.3.1). Perhaps the most prominent of these problems is David Hume's problem of induction. However, if one assumes that the inductive and abductive inference rules can be formulated as principles within a system of deductive rules only, then one can also embed most of these problems into the above framing of the problem of epistemic justification. The idea is to reduce all problems of higher level justification to a single problem of higher level justification, namely the problem of how to justify deduction, and an increasingly large set of lower level justification problems concerning the justification of basic propositions, namely the principles underlying inductive and abductive inferences. Since deductive inferences are truth preserving, their justification might be considered to be the least problematic one (we will discuss this justification problem briefly in chapter 8). Among a similar line of argumentation we could frame the problem differently such that it tends towards the other extreme: Given the (partial) intertranslatability of principles and rules, one can decrease the number of necessary lower level justifications (to a completely arbitrary degree) by reducing the set of basic propositions, while at the same time increasing the number of necessary higher level justifications by increasing the set of inference rules. Figure 1.2 visualises this fact which is well known from logical calculi (where there are systems with a couple of axioms and few inference rules and systems with no axioms at all, but many inference rules). In this sense, the foundationalist, coherentist, and infinitist approaches described above can be equally well considered as approaches to the problem of higher level justification.

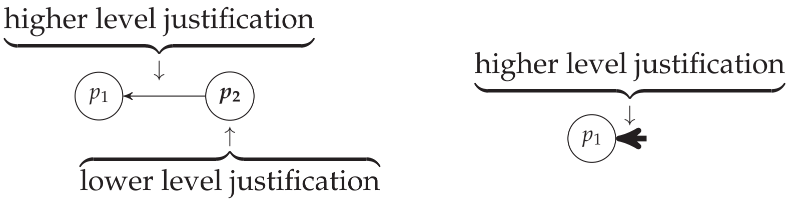
higher level justification
p_1
p_2
lower level justification
higher level justification
p_1

Figure 1.2: Mutual (partial) reducibility of the problem of higher level justification and lower level justification: Usually, the fewer inference rules a system has, the stronger its basis needs to be in order to allow for enough inferences. In this case one faces two epistemic problems of justification: a problem of higher level justification and several problems of lower level justification (left side). Typically, also the more inference rules a system has, the weaker its basis can be in order to account for the same inferences. In the extreme case the problem of lower level justification vanishes completely and one faces several problems of higher level justification (right side).

To sum up, the problem of epistemic justification consists in the fact that non-dogmatic, non-circular, and finitary constraints on justification and reasons are—taken together—only compatible with scepticism. To overcome the problem there are several alternatives on the market: One can give up at least one of these constraints and accept foundationalism, coherentism, or infinitism. Or one can bite the bullet and accept scepticism.

As the achievements of meta-induction show and as the further investigations in this book aim to show, is that scepticism can come in different forms: Being an epistemic sceptic does not imply that one needs to consider all notions of epistemic justification as empty. One can accept that the strict notion of epistemic justification as characterised by (EJ2)–(EJ4) is indeed not instantiated by any belief we have (i.e. one denies (EJ1)). However, instead of throwing out the baby with the bath water and claiming that none of our beliefs can be justified at all, one might try to consider other notions of epistemic justification and check whether they allow for rationalising our beliefs. An alternative on the market that is concerned with such a different notion of epistemic justification is naturalised epistemology. Here the idea is to accept the sceptical consequence of the strict notion of justification as characterised by (EJ2)–(EJ4), but at the same time to introduce a relative notion of justification which does not demand perfect or good enough epistemic performance, but simply *optimal* epistemic performance—optimal in the sense of being best compared to all the other available alternatives in the light of some given desiderata.

Here we will not discuss plain scepticism further. Rather, we will investigate and pick out the relevant parts of it in the context of methodological scepticism in section 2.2 (pp.51–54). In the remainder of this chapter we want to briefly discuss some merits and problems of the mentioned alternatives: foundationalism (section 1.1), coherentism (section 1.2), and infinitism (section 1.3). Afterwards, we roughly sketch the programme of naturalised epistemology and identify relevant parts of our approach therein (section 1.4).

1.1 Foundationalism

Foundationalism comes in different forms. There is classical foundationalism which claims that the "foundation of knowledge" is immediately intuitively graspable. A strong version of classical foundationalism also puts forward an infallibility constraint for the foundation. Modern versions of foundationalism are more moderate regarding the infallibility constraint and allow for a much weaker basis. Furthermore, the notion of justification approached by foundationalism is sometimes spelled out in internalist terms, i.e. justification is based on reasoning internal to the subject, i.e. reasoning the subject is aware of. This was the standard approach in almost all traditional foundationalist approaches. And sometimes it is spelled out in purely externalist terms, i.e. independent of some awareness of justifying reasons. This is an approach which arose especially in course of a turn in epistemological research caused by Gettier (1963).

Roughly speaking, all forms of foundationalism have at their basis the claim that we are justified (J) in believing something (EJ1), that justification is neither directly nor indirectly circular (EJ3), and that the chain of justification and reasoning comes to an end (EJ4). More specifically, foundationalism even states necessary and sufficient conditions for justification J. In Aristotle, e.g., we find such a definition in his characterisation of the axiomatic method where intuitively graspable principles make up for the axiomatic foundation of science. Thomas Aquinas provides a similar characterisation:

> "Now a truth is subject to a twofold consideration—as known in itself, and as known through another. What is known in itself, is as a 'principle,' and is at once understood by the intellect. [...] On the other hand, a truth which is known through another, is understood by the intellect, not at once, but by means of the reason's inquiry[.]" (see Thomas Aquinas 1981, answer to Quaestio 57, a2)

And in René Descartes we find it by his reference to the method of geometry, namely putting forward a set of axioms and making demonstrations which should serve as a general pattern for justification:

> "Those long chains of utterly simple and easy reasonings that geometers commonly use to arrive at their most difficult demonstrations had given me occasion to imagine that all the things that can fall within human knowledge follow from one another in the same way." (Descartes 1637/1998, part ii, p.11)

This classical foundationalist approach to justification was perpetuated by adherents of *rationalism* as, e.g., Gottfried Wilhelm Leibniz. (Note, since as a foundationalist Leibniz denied (EJ2), (EJ2) must not be mixed up with his *principle of sufficient reason* which is not an epistemic, but a metaphysical principle and according to which every effect has a cause or nothing *happens* without a reason.) Foundationalism was also the main approach to justification of the *empiricist* camp as, e.g., of Hume. The difference with respect to justification between these two epistemic schools can be nicely described by the schema of figure 1.2: Whereas rationalists started with strong metaphysical principles (a heavy basis), they allowed mainly deductive inferences. Empiricists, on the other side, started with a minimal basis (simple observations), and allowed for much stronger inferences, as, e.g., inductive inferences.

Not only rationalists and traditional empiricists, but also *logical positivists* and *logical empiricists* like Carnap in his *Aufbau* (1928) subscribed to foundationalism. However, they were already relaxing the infallibility condition for the basis towards a much more conventionalist stance. All these foundationalist approaches to justification are based on an underlying principle, which we call the 'main principle of foundationalism'. It is as follows:

Main Principle of Foundationalism.

(F) A belief is justified iff it is in our epistemic basis or it can be inferred from propositions of our epistemic basis.
Schematically: $Jx \leftrightarrow (Bx \vee \exists y(By \,\&\, yRx))$

Note that if B is non-empty and if we think of R as the classical deductive relation of *being logically strictly stronger*, then we can derive from the main principle of foundationalism (F) already (a very rough version of) non-scepticism (EJ2), non-circularity (EJ3), and finiteness (EJ4): This relation is transitive and asymmetrical; due to the compactness theorem of first order logic proofs of first order logic can be always reduced to finite ones; and some beliefs are justified, namely those in B. We can also deduce the claim that justification is inherited from reason to conclusion:

> If a proposition that is a reason for another proposition is justified, then also the other proposition is justified.
> Schematically: $\forall xy((Jy \,\&\, yRx) \rightarrow Jx)$

The principle of justification given up by foundationalism, namely (EJ2), stated the other direction: If something is justified, then there is a reason for it which is also justified.

As we mentioned above, classical foundationalism supplemented the main principle (F) with an assumption about being immediately evident and being infallible regarding the basis B: Every belief of B is immediately evident and infallible. For rationalists this evident and infallible basis consisted of fundamental epistemic and metaphysical principles like, e.g., Descartes' *Cogito* or Leibniz' *principle of sufficient reason*. For empiricists, the foundation consisted mainly of beliefs due to sense experiences and, according to our framing of the problem, a principle of induction. In this framing the higher level justification problem still consists of the justification of deductive reasoning which is, for most authors, taken for granted due to its property of truth-preservation. The problem of justifying the principle of induction as assumed by empiricists will be discussed in detail in part II of this book.

Similarly to traditionalist version of foundationalism, logical positivists and empiricists also supplemented the main principle (F) further: Although they did not subscribe to infallibilism of B, they still thought it to be most elementary in many respects (see "elementary experiences" in Carnap 1928/2003). However, what was really new was their expansion of the inference rules to logical, definitional, and mathematical inferences:

> "The method of logical analysis is what distinguishes the new empiricism and positivism from the earlier one, whose orientation was more biological-psychological." (Verein Ernst Mach 1996)

Although there are several problems with such an approach, we want to focus on what is often considered to be the main problem underlying this form of foundationalism—namely the *justification of the basis*: As perhaps most prominently Laurence BonJour pointed out, the choice of a basis is crucial to this approach, for which reason the question of how to justify such a choice comes up. The form of such a justified choice should be this: First, one shows that all beliefs in B have a specific feature ϕ—e.g., that they are granted by observation. Second, one shows that beliefs having feature ϕ are highly likely to be true. And finally, by this one concludes that all beliefs in B are highly likely to be true (see BonJour 1988, p.31). However, such a justification of choosing a specific B is not possible. E.g., regarding beliefs that are immediately evident: to justify them would mean that they are not immediately evident, but inferred. Usually this argument against the possibility of a justified choice of the foundation is put forward with respect to an empirical basis as follows:

> "**Basic Antifoundationalist Argument.**
>
> (1) Suppose that there are *basic empirical beliefs*, that is, empirical beliefs (a) which are epistemically justified, and (b) whose justification does not depend on that of any further empirical beliefs.
>
> (2) For a belief to be epistemically justified requires that there be a reason why it is likely to be true.
>
> (3) For a belief to be epistemically justified for a particular person requires that this person be himself in cognitive possession of such a reason.
>
> (4) The only way to be in cognitive possession of such a reason is to believe *with justification* the premisses from which it follows that the belief is likely to be true.
>
> (5) The premisses of such a justifying argument for an empirical belief cannot be entirely a priori; at least one such premise must be empirical.
>
> Therefore, the justification of a supposed basic empirical belief must depend on the justification of at least one other empirical belief, contradicting (1); it follows that there can be no basic empirical beliefs." (BonJour 1988)

Premise (1) results from denying (EJ2) of the problem of epistemic justification and putting forward the main principle of foundationalism (F). Premise (2) makes the assumption about justifying the choice of a basis B explicit, namely the assumption that such a choice needs to be backed up by features that make elements of B highly likely to be true. Premisses (3) and (4) are about the internal states of the epistemic agent making a choice regarding B. Finally, premise (5) is analytical and explicitly excludes, so to say, an *a priori-a posteriori fallacy*.

For a foundationalist clearly (1) and (5) hold, and also (2) once one accepts the task of justifying the choice of a basis B. So, the premisses that remain for a foundationalist to reject are (3) and (4). Here is a point where usually a prominent fission in approaching epistemic justification shows up: There is an *externalist response* to the basic antifoundationalist argument which denies (3). And there is an *internalist response* to this argument which denies (4).

Let us consider the *externalist response* first: A foundationalist may object that it is not necessary that the epistemic agent *has* a justification for her beliefs in B. Rather, what matters is that her beliefs in B are *de facto* caused in the right way, that they are *de facto* the product of a reliable belief forming process. What is very important is that she need not be aware of this. For achieving our aim of ending up with true beliefs *de facto* reliably forming beliefs is sufficient, there is no need for the epistemic agent to be aware of this. In this sense the justification of the choice of B is *external* to the agent, and in this sense denying premise (3) provides an externalist theory of epistemic justification J. A main proponent of this approach is, e.g., Goldman (1979) who argues for a reliabilistic theory of justification. One advantage of such a theory is that it allows for a nice solution of the problem influentially put forward by Edmund L. Gettier: In a Gettier case when an agent luckily forms a true belief, e.g., via deducing it from a justified, but false belief ('Jones owns a Ford' which is false but one might be justified in believing it, and 'Jones owns a Ford or Brown is in Barcelona' which is true and seems to be also justified since it is deduced from a justified belief), the agent's belief formation is not due to a reliable process since deduction from false premisses is clearly not reliable. Similarly for the case of *faked barns*. Although the agent *is not aware* that her true belief is unreliably formed due to the *daemonic environment*, it *de facto is* unreliably formed, and this is what counts for denying justification.

Clearly, this approach to the basic antifoundationalist argument comes at cost. A very strong argument against it is that the externalist concept of justification treats cases as different which are not distinguishable for an epistemic agent: E.g., in a fake barn setting the agent lacks justification, whereas in a real barn setting the agent has justification although both cases are indistinguishable for the agent. In this sense externalism seems to be relevant only for an outsider, someone who has a *God's eye view*. If we take for illustrative purposes the traditional analysis of *knowledge* as *justified true belief*, then *truth* clearly is a notion we can ascribe only from such a perspective. *Belief* is a notion we tend to completely "internalise" in the sense that whether it applies to some mental state or not is completely dependent on the respective agent in question. Finally, *justification* should serve as an intermediary between *truth* and *belief*, something which relevantly contributes to the difference between truly believing something and knowing something. This is also expressed in the so-called *value of knowledge problem* (see Pritchard 2007) or *Meno problem*:

> "Socrates: I will tell you. A man who knew the way to Larissa, or anywhere else you like, and went there and guided others would surely lead them well and correctly?—Certainly.
> Socrates: What if someone had had a correct opinion as to which was the way but had not gone there nor indeed had knowledge of it, would he not also lead correctly?—Certainly."
> (Plato 1997, p.895, Meno, 97c)

If J has no internalist component at all, it seems hard to argue for a difference between knowing something and simply truly believing something. And it seems that we aim at ending up *knowing* the principles of epistemic justification and not just *truly believing* them.

Let us come to an *internalist response* to the basic antifoundationalist argument, namely by objecting (4): According to this version of foundationalism, the epistemic agent is in possession of reasons for choosing B, but this does not mean that she believes with justification that all beliefs in B are likely to be true. Rather, a basic belief is an immediate awareness, a self-evident intuition, which needs no further justification. This is also called *the idea of the given* (see Pojman 2000, p.110). A main proponent of this approach is, e.g., Roderick M Chisholm (1989), however, all classical foundationalists as, e.g., Descartes, Leibniz, John Locke, and Hume, and also later on almost all foundationalists before the Gettier turn in epistemology are regarded as internalists in this respect.

Again, there is also a payoff for the internalist approach. Very prominent is, e.g., Wilfrid Sellars critique of the *myth of the given* (see Sellars 1991, chpt.5: Empiricism and the Philosophy of Mind) which basically states that if we refer to immediate awareness etc. we provide no justification for the choice of B. And indeed, if one denies (4) and refers to self-evident intuition, then one no longer provides justification of the choice of B in terms of reasons that make the beliefs in B likely to be true: Prima facie it is questionable whether *truth* (in terms of *God's eye view*) and *self-evidence* (the individual perspective) are related in this way. Furthermore, it is questionable whether principles needed for justifying a bulk of our knowledge such as the principle of induction can be considered to be self-evident. And if one were to avoid this problem by considering the principle in terms of an inference rule, then she would just shift the burden of proof from the lower level problem of justification to a higher level justification.

In general we want to conclude that the *classical* as well as modern *internalist foundationalist* approaches to epistemic justification face a dilemma: If only beliefs are in the basis B that are immediately evident, then either the problem of scepticism or the problem of higher level justification shows up again: The former is the case if we keep the inferences fixed (just deductive ones), but have to throw out many beliefs from our basis B due to their not being self-evident. Such restrictions regarding B and inferences do not allow for justifying enough beliefs. The latter is the case if we allow for the justification of enough beliefs by shifting non self-evident beliefs of the basis to corresponding inference rules: In such a case, one needs to provide higher level justifications which a foundationalist approach to justification in the strict sense cannot provide: Truth-preservation is characteristic and guaranteed for deductive inferences only, but not for any other. Finally, *externalist foundationalist* approaches to justification face the problem that they provide a theory of justification which seems to be relevant only for epistemic agents that can take in a *God's eye view*. For such a position the relevance of epistemic justification seems to vanish since the value of knowledge in the sense of justified true belief and that of true belief simpliciter also seems to vanish: If—for deciding whether a belief is justified or not—an epistemic agent has to take in a *God's eye view*, why then not using *God's eye view* directly for deciding whether a belief is true or not?

In the next sections we are going to briefly discuss foundationalism's main rivals, i.e. coherentism and infinitism, before we go on with outlining our approach of *epistemic engineering*.

1.2 Coherentism

Coherentism is an approach to epistemic justification according to which coherence is not only a necessary condition for justification, but also a sufficient one. Many traditional positions as, e.g., Plato's or Georg W. F. Hegel's theories of truth have been interpreted as coherentist accounts (see Pojman 2000, p.116). However, it is important to note that most traditional accounts were concerned with a coherentist understanding of *truth*. Most modern coherentists like Quine, Sellars, and BonJour reject the coherence theory of *truth*, but support a coherentist theory of *justification*:

> "The indicated conclusion is that there is no real alternative to the standard and commonsensical conception of truth as, roughly, correspondence or agreement with independent reality; and thus that a satisfactory metajustification for our envisaged coherentist theory of empirical justification must involve showing in some way that achieving coherence in one's system of beliefs is also at least likely to yield correspondence." (BonJour 1988, p.158)

Roughly speaking, coherentism has as its basis the claim that we are justified (J) in believing something (EJ1), that we can provide justified reasons for all justified beliefs (EJ2), and that the chain of justification and reasoning comes to an end (EJ4). What coherentism denies is that justification is never directly or indirectly circular (EJ3). Perhaps the most colourful metaphor used in this context is that one of Otto Neurath which was popularised by Quine (1963a, p.79): “The philosopher’s task was well compared by Neurath to that of a mariner who must rebuild his ship on the open sea.” Similar metaphors can be also found in explanations of *bootstrapping arguments* (originally this term was used in computer science for describing self-starting processes which were supposed to proceed without external intervention), in John D. Norton’s (2014-03) arches where the ultimate support for each stone derives from many stones both above it and below it and then from the entirety of the stones in either side of the arch, or also in literature as, e.g., *Baron Muenchausen* who pulls himself out of a swamp by his own hair (backed up on this story Hans Albert called the problem of epistemic justification also the *Muenchhausen Trilemma*).

Coherentism is mainly motivated by the problem of the epistemic basis B: As we have seen in the preceding section, foundationalism either lacks a justification of the choice of B, or, if one can provide a convincing justification for the choice of B, then she must source out many elements of B to the inferences and lacks justification there. So, in general it seems that any consensual choice of B (as, e.g., elementary experiences, protocol sentences, observational statements etc.) and inference rules (as, e.g., deductive ones) does not allow for justifying principles we typically consider to be justified. In this sense, B (and the inference rules) underdetermine justification J. This led, e.g., Quine to a holistic position regarding justification according to which nothing is justified *per se* or *absolutely*, but only in the context of a *web of belief* (see Quine and Ullian 1978). According to this position J is even that much underdetermined by B that one might even revise logical inferences (see Quine 1963b, sect.6).

It is common to distinguish two forms of coherentism: There is *linear coherentism* which plainly assumes that there are chains of justification which are directly or indirectly circular (as illustrated in figure 1.1). And there is *holistic coherentism* or *emergence coherentism* which seems to not state direct or indirect circularity, but to consider justification as a holistic concept (see Elgin 2014; and Pojman 2000, p.116). However, in order to present the main critique against coherentism and in order to remain in the framework from above we do not need to make this distinction, because the relevant and problematic properties of the *holistic notion of justification* can be equally well expressed by help of the *linear notion*. The relevant bridge principle is that if we are holistically justified in a set of beliefs, then we are also justified in using single beliefs in our linear reasoning chains. Or, in more figurative words, the holistic property of *coherence* grants justification in linear reasoning. Taking this into account, we can formulate the main principle of coherentism, which supplements (EJ1), (EJ2) and (EJ4), as follows:

Main Principle of Coherentism.

(C) A belief is justified iff its respective system of beliefs is coherent.

> Schematically: $Jx \leftrightarrow C(\iota y)(yBx)$

Here $(\iota y)(yBx)$ is intended to single out *the* overarching belief system y of x—yBx stands for: y is an overarching belief system of x; below we will further comment on the assumption that for any x there is exactly one such overarching belief system. Note that the role of the basis B for justifying a belief in a proposition p in the foundationalist picture is now taken over by a coherent belief system $b = (\iota y)(yBp)$. Note also that from (C) we get the implausible consequence that all beliefs of one and the same belief system are either equally justified or unjustified (this is similar to the general inheritance principle for justification along a line of reasoning which holds for foundationalism—which is in some sense more general, because according to foundationalism all beliefs are part of one belief system, so to speak, namely that one which is generated by help of the basis and the inference rules):

> Two beliefs of one and the same belief system are equally justified or unjustified.
> Schematically: $\forall xy((\iota z)(zBx) = (\iota z)(zBy) \rightarrow (Jy \leftrightarrow Jx))$

This holds, because if, e.g., the belief system of x and y is b, then via (C) x's justification (Jx) implies b's coherence (Cb) which in turn, again via (C), implies y's justification (Jy) and vice versa.

Clearly, there are several parts of principle (C) which are in need of further specification: First of all, what is the respective system y of a belief x (yBx) and is there *exactly one* such system for any belief? The answer is: "No". However, we want to make this "modeller assumption", because we need not discuss problems related to one and the same belief being part of different belief systems. As we will see below, the main concerns with coherentism show up already in this simplified model. For this reason we want to focus on the second pressing point which needs to be specified, namely what does it mean that a system of beliefs is coherent?

The notion of *coherence* C is used quite differently in the literature. If we take a set of propositions $b = \{p_1, p_2, \ldots\}$ to be a system of beliefs, then a quite elementary proposal for considering b to be coherent (Cb) is to demand logical consistency of b and deducibility within b in the sense that every proposition p_i follows from $b \setminus \{p_i\}$ (this was, e.g., the position of Alfred C. Ewing as described in Olsson 2018, sect.3). This is a very strong notion of coherence which is satisfied only by very little belief systems as, e.g., $\{p_1, p_2, p_1\&p_2\}$, but, e.g., not by $\{p_1, p_2, p_1 \rightarrow p_2\}$.

By help of this very strong notion of *coherence* which makes up for a very weak notion of *coherentism* we can already illustrate the circularity of coherentist accounts of justification and reasoning: If such an account supplements an account of reasoning as, e.g., characterised by (EJ1), (EJ2) and (EJ4), then it follows immediately from the incompatibility of (EJ1), (EJ2) and (EJ4) with (EJ3) that such an account is directly or indirectly circular. Clearly, the principle of coherentism (C) does not need to be considered as supplementing an account of reasoning satisfying the mentioned desiderata. But if it does, then it is circular. We can illustrate this by considering a system of beliefs $b = \{p_1, p_2, p_3\}$, where $p_3 = `p_1\&p_2'$. As we have seen above, b is coherent. Hence, by (C) we get Jp_1, Jp_2, and Jp_3. For simplicity reasons let us assume that this is the only coherent system for p_1, p_2, p_3 (there are others, but, of course, circularity follows for them too). By (EJ2), the principle which connects justification J with reasoning R, we get that for p_1, p_2, and p_3 there must be a justified reason: $\exists x Jx \;\&\; xRp_i$. Since exactly p_1, p_2, and p_3 are justified, this means that either at least one of the p_i's is a reason for itself, i.e. there is a direct reasoning circle, or at least one of the p_i's stands to itself in the relation of the transitive closure of R, i.e. there is an indirect reasoning circle (e.g.: p_1Rp_2, p_2Rp_3, p_3Rp_1). An even simpler example is $b = \{p_1\}$ which is "vacuously" coherent and enforces a direct reasoning circle.

Let us come back to the characterisation of C: Already more powerful regarding justification is CI Lewis' proposal of interpreting C in the sense of probabilistic dependency: Cb iff $Pr(p_i|p_{1\neq i}, \ldots, p_{n\neq i}) > p_i$, where $p_i, p_1, \ldots, p_n$ are all the propositions of b (see "congruence" in Lewis 1946). Since (almost) all bs that satisfy the Ewing-criterion for C also satisfy the Lewis-criterion, Lewis' notion of *coherence* is wider. One could also think of relaxing the criterion by, e.g., not demanding probabilistic increase of p_i conditional on the remainder set, but such increase of p_i conditional on subsets of the remainder etc. (see Olsson 2018, sect.3; for a general overview of common probabilistic coherence measures see Hartmann and Sprenger 2011, sect.4).

Even more moderate is the approach of BonJour who stated some desiderata for a quantitative notion of *coherence* (BonJour 1988, pp.95-99):

Some Desiderata for Coherence.

(1) "A system of beliefs is coherent only if it is logically consistent.

(2) A system of beliefs is coherent in proportion to its degree of probabilistic [coherence].

(3) The coherence of a system of beliefs is increased by the presence of inferential connections between its component beliefs and increased in proportion to the number and strength of such connections.

(4) The coherence of a system of beliefs is diminished to the extent to which it is divided into subsystems of beliefs which are relatively unconnected to each other by inferential connections."

Desiderata (1) and (2) cover the consistency constraint of Ewing as well as a probabilistic version thereof (b does not contain any p_i and $p_j = $ '$Pr(p_i) < r$', where r is some threshold < 0.5). Desideratum (3) captures Ewing's deducibility constraint as well as Lewis' constraint of positive correlation. Desideratum (4) considers such interconnections in a more fine-grained way (as indicated above by not only considering correlations conditional on the remainder set, but also pairwise positive correlations conditional on subsets of the remainder etc.). Putting forward several desiderata for C opens up the problem of balancing between them. This is, however, not the concern we have. The main concern we have is that in whatever way C is characterised, for a solution to the problem of epistemic justification one needs to provide an argument that C is *truth conducive*. Otherwise the choice of C faces the same problem as the choice of B faces for the fundamentalist (recall, the problem was that, e.g., *self-evidence* needs to be truth conducive in order to provide a justification for a self-evidence-based choice of B). Qualitatively speaking, one needs to show that if Cb_1 whereas $\neg Cb_2$, then b_1 is more likely to be true than b_2 ($Pr(b_1) > Pr(b_2)$). Quantitatively speaking, this means that one needs to show that given the degree of coherence of b_1 is greater than that of b_2, then also $Pr(b_1) > Pr(b_2)$. However, there are impossibility results which show that truth conduciveness cannot be guaranteed for any characterisation of (a measure of) C which satisfies informational constraints in the line as mentioned above (see, e.g., Bovens and Hartmann 2003, sect.1.4).

Let us illustrate this by help of a simple example: One popular case that is often put forward in order to argue for the *emergence of justification* out of coherence is the case of consilience of surprising testimony. We will discuss this case in detail in chapter 9. Here we focus just on some very simplified features: Assume that h is an event very unlikely to occur. And assume that there are n testifiers that are independent of each other, but all testify h: $e_1, \ldots, e_n$. Since they are independent of each other, it seems that the consilience of the surprising testimonies makes h more likely to be true, at least more than considering each single testimony. I.e.: $Pr(h|e_1, \ldots, e_n) > Pr(h|e_i)$ for all $1 \leq i \leq n$. So, it seems that the coherence among the surprising testimonies allows for *emergence* of justification:

> "Evidently the best explanation of the agreement is that the reports are true. [...] The thesis of the sort of epistemological holism that I want to consider is that epistemic justification is primarily a property of a suitably comprehensive, coherent account, when the best explanation of coherence is that the account is at least roughly true." (see Elgin 2014, pp.245f)

However, given the independence of the single testimonies $e_1, \ldots, e_2$ and by applying Bayes' theorem it turns out that this assumption is equivalent to simply stating that all agents are positive truth trackers regarding

h: $Pr(e_i|h) > Pr(e_i)$. So, the alleged *emergence* of justification consists de facto in *presupposing* justification. Something similar holds if one stresses so-called *wise crowd* and *Condorcet* effects for coherentism—we will discuss them in detail in chapter 12. In order to achieve these effects one also has to make independence and reliability assumptions before one can cash them out for justification.

Missing truth conduciveness is not the only problem with coherentism. Similarly as is the case for foundationalism, also an internalist version of coherentism seems to be prone to a regress-problem:

> "If we embrace access [i.e. an internalist version of] coherentism, then coherentists face the very regress that traditional foundationalists tried so desperately to avoid. To justifiably believe that our beliefs cohere we would need to know first what we believe and second that the propositions believed stand in the appropriate evidential relations. But as coherentists we have no foundations to fall back on. We can't just give ourselves privileged access to propositions describing our own belief states. Our only access to what we believe is through a coherence we discover between our belief that we have certain beliefs and the rest of what we believe. But to discover this coherence we will once again be forced to discover what we believe, and so on, ad infinitum." (Fumerton 2002, pp.229f)

As Pojman (2000, p.120) puts it, the problem for internalist coherentism is that the belief system of an agent needs to be coherent in order to be justified (C). However, in the internalist version this means that an agent needs to be aware that her system is coherent. But since there is no coherence-basis one could stop at, this leads to the following infinite regress:

p_1: My belief set is b and p coheres with b.

p_2: My belief set is b and p_1 coheres with b, i.e.:
My belief set is b and that p coheres with b coheres with b.

p_3: My belief set is b and p_2 coheres with b.

$$\vdots$$

Let us make the regress-problem explicit: Given that every belief is part of exactly one belief system, it seems to be plausible to assume that the belief system of a belief system (considered as a belief) and the belief system (considered as a belief) itself are identical with each other, i.e. $b = (\iota y)(yBb)$. But then, by (C), every coherent belief system is justified ($Jb \leftrightarrow Cb$, we will discuss a problem with this justificatorial inflationism immediately). Now, according to an externalist coherentist conception of justification this seems to be perfectly fine: Once we can decide, from a *God's eye view* whether b is coherent (Cb), we can also ascribe justification to b. However, for an internalist coherentist an answer to the question whether Cb or not needs to be accessible or *internal* to the epistemic agent. Hence, for an internalist coherentist (C) needs to be modified to $Jx \leftrightarrow JC(\iota y)(yBx)$. Hence, she demands JCb for Jb and not simply Cb as the externalist does. And this is where the regress starts: In order to justify p (Jp), we need to justify that its belief system b is coherent (JCb), which in turn requires us to justify that Cb's belief system, let us say b', is coherent (JCb'), and so forth.

Even if these problems could be overcome and there were *emergence* or *bootstrap* justification, coherentism would have the problem that it grants too much justification in the following sense: Given finite belief systems b_1, b_2 and Cb_1 and Cb_2 by the main principle of coherentism (C) we get Jb_1 and Jb_2. Now, it is not hard to think of examples where b_1 and b_2 are incompatible with each other. Even if we take the very strict notion of Ewing from above, we get, e.g., $b_1 = \{p_1, p_2, p_1\&p_2\}$ as well as $b_2 = \{\neg p_1, \neg p_2, \neg p_1\&\neg p_2\}$ to be coherent, although b_1 and b_2 are logically inconsistent. Note that b_1 and b_2 share no belief, which shows that this problem shows up already with our strict "modelling assumption" that every belief is part of exactly one belief system. This argument amounts to a self-refutation argument with the conclusion that coherentism is incoherent. A typical coherentist response consists in claiming that not all coherent systems allow for justification, but only compatible ones: So, either b_1 or b_2 is justified, but not both. However, this objection to the self-refutation argument brings in the problem we were initially concerned with, namely the underdetermination of J—this time not with respect to the basis B and the inference rules, but with respect to (C). This problem will show up again when we discuss a coherentist approach to Hume's problem of induction in chapter 5.

At first glance, coherentism seemed to be a promising approach to the problem of epistemic justification with respect to the underdetermination of J given B and inference rules we justifiably use in our reasoning. However, it seems that the justification of coherentism in terms of truth conduciveness fails, that coherentism in its internalist variant faces a regress problem, and that the problem of underdetermination pops up again. *So far, so bad* for coherentism. But what about accepting infinite justificatorial regresses? Is this a viable route?

1.3 Infinitism

Aristotle already discussed infinitism in his *Posterior Analytics*—the heading of chapter 3 of book I is *Two errors—the view that knowledge is impossible* *because it involves an infinite regress, and the view that circular demonstration is satisfactory* (see Aristotle 1957, 72b, p.512). He clearly objected to it and mentioned as one important reason our finite reasoning capacities. And in fact, it seems that all non-sceptical traditional approaches afterwards did not even consider infinitism a viable option. Hume wrote, e.g.:

> “If I ask, why you believe any particular matter of fact, which you relate, you must tell me some reason; and this reason will be some other fact, connected with it. But as you cannot proceed after this manner, *in infinitum*, you must at last terminate in some fact, which is present to your memory or senses; or must allow that your belief is entirely without foundation.” (Hume 1748/2007, p.33)

In the last sentence it seems the he speaks of *justification* when he uses ‘*foundation*’. This equating of *being justified* and *having a foundation* seems to be symptomatic for the predominant finitist approaches to justification.

Whereas finitist approaches stress the importance of a foundationalist or coherentist basis B for justification, infinitists stress the importance of providing reasons (R) for justification. On this account they think that reasoning chains never stop or loop: Halting is prevented by demanding to always provide a reason for a justified belief. And looping is prevented by demanding *new* reasons for such a belief. These are the reasoning constraints on justification which an infinitist embraces. So, infinitism has as its basis the claim that we are justified (J) in believing something (EJ1), that we can provide justified reasons for all justified beliefs (EJ2), and that reasoning is neither directly nor indirectly circular (EJ3). What infinitism denies is that justification ever comes to an end (EJ4). So, the main principle of infinitism supplements (EJ1) and consists of the following strengthening of (EJ2) and (EJ3):

Main Principle of Infinitism.

(I) A belief is justified iff it has a justified reason, where reasoning is never directly or indirectly circular.
Schematically: $\forall x(Jx \leftrightarrow \exists y(Jy \,\&\, yRx))$ &
$\forall xyz(xRy \rightarrow (\neg yRx \,\&\, yRz \rightarrow xRz))$

As an early proponent of infinitism counts Charles S. Peirce (see Aikin 2011, sect.3.2). Contemporary proponents are Klein (1998), Fantl (2003), Aikin (2011), and Atkinson and Peijnenburg (2009). Besides directly incorporating the above intuitions regarding the role of reasons in justification, infinitism allows also for satisfying the so-called *degree requirement* for justification: The basic intuition behind this requirement is the idea that the more justified reasons one can provide for a belief, the better the proposition itself is justified. So, speaking in quantitative terms, the degree of justification of a belief should increase with the number of reasons provided for the belief (in a chain-wise manner). In the infinitist case "warrant increases not because we are getting closer to a basic proposition but rather because we are getting further from the questioned proposition" (see Klein 2014, p.280). And since the longer the reasoning chain, the more reasons are provided, in the infinitist case the degree of justification is supposed to asymptotically approach complete justification. To illustrate this, recall the reasoning chain of an infinitist from figure 1.1—it is as follows (since we are going to provide a causal model later on, we speak of hypotheses h_i instead of propositions p_i now):

$$h_1 \xleftarrow{R} h_2 \xleftarrow{R} h_3 \xleftarrow{R} \cdots$$

Now, if, e.g., h_1 is the belief that Jones is the murderer, then providing as reason the belief h_2 that he had a knife seems to increase justification of h_1. If one can provide, e.g., as a further reason for h_2 the belief h_3 that a knife is missing, this seems to increase justification (not only for h_2, but also) for h_1 further, and so forth. So, infinitism seems to allow for the *emergence of justification* by simply continuing reasoning.

It is difficult to make exact technical sense of this. Here is a suggestion which is also intended to show that this intuition behind infinitism is suspicious: Assume the reasoning chain of above. And assume further (this is expanding the chain model a little bit), that each belief h_i used as a reason is supplemented with some evidence e_i: So, e.g., for h_2 (my belief that Jones had a knife) we have evidence e_2 that I have seen Jones with a knife, and for h_3 (my belief that a knife is missing) we have evidence e_3 about me counting the knives before and after the murder etc. Now, in this case these reasons and evidential reasons seem to be related as follows: The h_is and e_is are positively correlated (my seeing Jones with a knife increases the probability of (me believing of) him having a knife etc.). Furthermore, each h_{i+1} seems to *screen off* h_i from each further reason $h_{>i+1}$ in the chain in the following sense: If I get to know or if I am fully justified in h_{i+1}, then there is no probabilistic increase in h_i once I also got to know $h_{>i+1}$, since the burden of justification of h_i is mediated completely via h_{i+1} and h_{i+1} is considered to be already fully justified. So, e.g., if I know that Jones had a knife, then my belief that a knife is missing brings about no further justificatory support for me believing that Jones was the murderer. Finally, it seems to be also perfectly reasonable to assume that the correlations between the reasons and the evidential reasons are not strict and that my reasons (beliefs) are not fully justified. Speaking in terms of probabilities, this means that we can represent the case by help of a so-called Bayesian network as depicted in figure 1.3 (for an excellent introduction to Bayesian networks see Gebharter 2017, sect.2 and 3; and Sprenger and Hartmann 2019, some basics of this framework are presented in section 2.1). Due to the mentioned *screening off*-property, this network represents the following probabilistic independencies:

$$\begin{aligned} Pr(e_i|h_i, X) &= Pr(e_i|h_i) \\ Pr(h_i|h_{i+1}, Y) &= Pr(h_i|h_{i+1}) \\ \text{where } X \subseteq \{h_j : j \leq n\} &\cup \{e_j : j \leq n\}, \text{ and} \\ \text{where } Y \subseteq \{h_j : n \geq j \geq i+1\} &\cup \{e_j : n \geq j \leq i+1\}. \end{aligned}$$

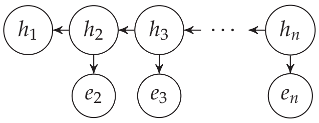

Figure 1.3: Bayesian network of a reasoning chain with evidential support: h_1 is the belief which is mainly reasoned for by help of a chain of reason-evidence pairs $\langle h_i, e_i \rangle$ which are positively correlated ($e_i \leftarrow h_i$ stands for a causal relation manifested in probabilistic increase: $Pr(e_i|h_i) > Pr(e_i)$; similarly for $h_i \leftarrow h_{i+1}$); also each reasoning pair $\langle h_i, h_{i+1} \rangle$ is positively correlated; given the above described independence assumptions, in the finite case h_1's justification in terms of probabilities grows with a growing number of such reason-evidence pairs $\langle h_i, e_i \rangle$.

Now, if we assume, e.g., that the correlations are all equally strong, i.e.: $Pr(h_i|h_{i+1}) - Pr(h_i) = Pr(h_j|h_{j+1}) - Pr(h_j) = Pr(h_k|e_k) - Pr(h_k) > 0$ (this also excludes that the probabilities of the hs are extreme), then it follows that also all es are positively correlated with h_1. Even more, the more evidential reasons (es) one provides, the higher the probability of h_1, i.e. the more h_1 is justified: It follows from the assumptions about correlations and independencies:

$$\text{If } j > i, \text{ then } Pr(h_1|\{e_k : 2 \leq k \leq j\}) > Pr(h_1|\{e_k : 2 \leq k \leq i\})$$

So, providing further reasons increases justification in our model. However, note that this holds only for finitely many (n) evidential reasons. For the infinite case, the situation is different. From the above assumptions it also follows that the justificatory impact of evidence decreases with distance: The difference between the probabilistic boost of h_1 by $e_1, \ldots, e_{k+1}$ and the probabilistic boost of h_1 by $e_1, \ldots, e_k$ shrinks with increasing k:

$$\begin{aligned} \text{If } j > i, \text{ then } & Pr(h_1|\{e_k : 2 \leq k \leq j\}) - Pr(h_1|\{e_k : 2 \leq k \leq j-1\}) \;<\; \\ & Pr(h_1|\{e_k : 2 \leq k \leq i\}) - Pr(h_1|\{e_k : 2 \leq k \leq i-1\}) \end{aligned}$$

Since with an increasing number of evidential reasons their probabilistic boost shrinks and vanishes in the end, at some point in the reasoning chain it will be epistemically insignificant whether one provides another reason or not. So, infinitism seems to provide no surplus justification to foundationalist justification.

With respect to the reasoning aspect of infinitism, Turri (2009) has argued for a foundationalist approach to epistemic justification which also allows for infinitely long reasoning chains. This clearly deviates from our terminology, since we defined 'foundationalism' as a position which excludes infinite reasoning chains. According to the understanding of Turri (2009), foundationalism is characterised by a basis B which is intended to justify all beliefs (this is in agreement with our characterisation), but that the justification or reasoning procedure never stops, because this would amount to dogmatism (this is different from our characterisation). He argues for a view "that endorses neither circular reasoning nor arbitrariness" and can be schematically characterised as follows (see Turri 2009, p.161):

$$h_1 \xleftarrow{R} h_2 \xleftarrow{R} h_3 \xleftarrow{R} \cdots \xleftarrow{R} B$$

The idea is that all beliefs can be justified on the basis of B, but that reasoning need not come to a dogmatic end. The example he provides is as follows: Assume that an epistemic agent has observed that it is 2:05. Since she observed it, it is part of her epistemic justification and reasoning basis B. Now, based on this observation she forms the belief and claim h_1 that it is past 2:00 o'clock (h_1:>2:00). Now, she can justify and reason for this claim, e.g., by help of her belief h_2: >2:02:30; again, she can justify and reason for this claim by help of her belief h_3: >2:03:45; in principle she could reason this way infinitely long by getting halfway closer to 2:05. By "proceeding this way ensures that [an epistemic agent] will approach the limit of, but never arrive at, 2:05. In other words, she has available to her an infinite series of non-repeating reasons, each of which is entailed by its successor. Moreover, the foundationalist has a principled story to tell about how each member of this infinite series gets justified for her: namely, she can see that it is 2:05" (see Turri 2009, p.163). This is an example according to which a foundationalist can be non-dogmatic in the sense that she can always provide a justified reason, in principle infinitely many of them. Hence, also with respect to the reasoning aspect, infinitism seems to not provide a surplus to foundationalism.

Furthermore, the *finite mind objection* (as approached, e.g., in Klein and Turri 2012, sect.2) still strikes many epistemologists as valid:

> "The infinite regresses are mushrooming out in an infinite number of different directions. If finite minds should worry about the possibility of completing one infinitely long chain of reasoning, they should be downright depressed about the possibility of completing an infinite number of infinitely long chains of reasoning. I call this the *epistemic* regress argument for foundationalism" (Fumerton 1995, p.57)

To sum up, infinitism clearly is an alternative approach to the problem of epistemic justification which deserves further investigation. However, it is conceptually hard for many epistemologists to get their head around a notion of *justification* that makes use of infinitely many reasons (perhaps a weaker notion of *justifiable, in principle* is better captured by it). Regarding the promise of an *emergence of justification* as well as regarding its non-dogmaticity it seems to be not better off than foundationalism, because it also demands to always provide non-circular reasons.

After travelling along the main routes to epistemic justification with their potholes and dead ends, it seems that we need to get on a new route. Let us try epistemic engineering which can be motivated excellently by taking the highway of naturalised epistemology.

1.4 From Naturalised Epistemology to Epistemic Engineering

In 1968, Quine gave a lecture on "Epistemology Naturalized" as an invited address to the *Fourteenth International Congress of Philosophy* in Vienna. (Large parts were adapted from the lecture "Stimulus and Meaning" he gave already in 1965 at Michigan State University.) According to followers of his programme, with this lecture he sounded the death knell of classical epistemology concerned mainly with the normative notion of justification. As a new start, so Quine, a new way of performing epistemological research is in need:

> "The Cartesian quest for certainty had been the remote motivation of epistemology, both on its conceptual and its doctrinal side; but that quest was seen as a lost cause. [p.74]
> Epistemology still goes on, though in a new setting and a clarified status. Epistemology, or something like it, simply falls into place as a chapter of psychology and hence of natural science. [p.82]" (Quine 1969)

This kind of naturalising epistemology is sometimes also called *replacement naturalism* (see Kornblith 2002) or *methodological naturalism* (see Goldman 1994, p.309). As it is often the case with new and very ambitious programmes, it seems that also here, at the early stage of naturalised epistemology, the baby was thrown out with the bath water: Drawbacks in finding a convincing account to the problem of the (*normative*) notion of epistemic justification led its adherents to fully abandon the enterprise of *normative* epistemology and focus on *descriptive* epistemology instead.

Contemporary proponents of the programme as, e.g., Hilary Kornblith are much more sophisticated and cautious. In his (1999), Kornblith starts with traditional *Cartesian epistemology* and characterises it as follows: Descartes proposed that a belief counts as knowledge iff it is foundational or derived from what is foundational. In such a foundationalist approach, even in the light of skeptical challenges, knowledge is possible if the foundation is immune from error. Since particularly empirical knowledge is prone to error, Descartes conceived the theory of knowledge as prior to empirical knowledge. This is the main pillar of the classical approach which Kornblith thinks that naturalistic epistemologists strongly disagree with: The naturalistic alternative considers epistemology not prior to, but *continuous* with the empirical sciences. A consequence of this is that knowledge is treated as a natural phenomenon which should be investigated also by empirical means. For example, investigating the social factors which give rise to paradigm cases of knowledge might be prior to building a theory of knowledge (see Kornblith 1999, p.163).

The main issue between traditional epistemology and naturalised epistemology is whether there are any epistemological investigations that are adequately performed in a purely a priori (non-empirical) way. Kornblith denies this, whereas proponents of the traditional approach take on this position. So, e.g. Feldman (1999) argues that scepticism is not just a *red herring* (see Pojman 2000, p.186), but of real concern and needs to be argued against on a purely a priori ground. Further objections to naturalised epistemology are the accusation of being *viciously circular* due to relying on empirical science as a legitimate source of knowledge, or being *self defeating* in the sense that to evaluate arguments in favour of epistemic naturalism seems to presuppose the "legitimacy of appeals to *a priori* or 'armchair' intuition" (see Rysiew 2018, sect.3.1). However, the perhaps most important objection to naturalised epistemology results from concerns with epistemic normativity. Once epistemology is considered to be a branch of descriptive sciences like psychology, it is descriptive itself and the question remains of how to deal with problems of epistemic normativity: We are not only interested in the question of how humans reason in reality, what are their de facto sources of knowledge, how they incorporate evidence etc. We are also concerned with normative questions of how they *should* do so, how an *ideal epistemic agent* would act etc. But how can a descriptive discipline like naturalised epistemology provide answers to such normative questions? As already Kornblith (2002, chpt.5) mentioned, after initiating the programme of naturalised epistemology, Quine was many times and at many occasions concerned with clarifying his account of naturalistic epistemology. In particular, this meant that he tried to clarify the role of normativity within epistemological theorising. A key passage is the following one:

> "For me normative epistemology is a branch of engineering. It is the technology of truth-seeking, or, in a more cautiously epistemological term, prediction [...]. There is no question here of ultimate value, as in morals; it is a matter of efficacy for an ulterior end, truth or prediction. The normative here, as elsewhere in engineering, becomes descriptive when the terminal param-

> eter is expressed." (see Quine 1998, pp.664f)

The idea is that based on a *means-end principle* of the form (see Schurz 1997, p.128, note that in Schurz' investigation the principle is about necessary means, whereas here we use the principle for any adequate means in a wider, but still absolute, sense):

> **Epistemic Means-End Principle.**
> If A is an epistemic end and if B is a necessary or adequate means for A, then also B is an epistemic end.
> Schematically: $(\mathcal{O}A \;\&\; (\Box(A \to B) \vee \Box(B \to A))) \to \mathcal{O}B$

one can argue for the normativity of some epistemic notions given some epistemic ends as, e.g., truth, prediction etc. In this relative harmless sense also Alvin I. Goldman can be considered as a prominent proponent of naturalised epistemology: "Science can help identify the forms of social organisation that would optimize the chances of obtaining epistemic ends." (see Goldman 1994, p.308).

Note that the question of the ulterior end(s) $\mathcal{O}A$ is not accounted for up to now. Naturalised epistemology suggests to also receive the *terminal parameters* from science. So, e.g., Kornblith (2002, chpt.5) identifies them with desires. However, if achieved this way, *normativity* would still consist in some natural property, a view which is highly non-consensual. For this reason we want to remain completely open in this respect. One might end up with such ends in a purely descriptive way. But one might also end up with such ends by purely normative and a priori considerations. What we want to take over from epistemic naturalism for our enterprise is the engineering part, the search for necessary ($\Box(A \to B)$) and adequate conditions ($\Box(B \to A)$) given some end A. This is what we consider to be the main task of *epistemic engineering*. The schema in figure 1.4 demarcates the *epistemic engineering* part from other parts of epistemic investigations.

$$\overbrace{(\mathcal{O}A}^{\text{ulterior epistemic ends}} \qquad \& \quad \underbrace{\Box(A \to B) \vee \Box(B \to A))}_{\text{epistemic engineering}} \quad \to \qquad \overbrace{\mathcal{O}B}^{\text{derived epistemic ends}}$$

Figure 1.4: The role of epistemic engineering in epistemic enquiry: Given some epistemic ends (A) it seeks necessary and adequate means (B) to achieve these ends, which allows also for deriving normative statements about the means.

Several remarks are in place: First of all, what makes a means a *necessary or adequate* means? We will employ the epistemic means-end principle only for two kinds of cases namely (i) where 'necessary' is understood in the sense of '*analytic*' and (ii) where 'adequate' means '*optimal*'. Regarding the former (i), $\Box(A \rightarrow B)$ means that from the principles of logic, mathematics together with some definitions one can derive B from A. This is a very strong case for deriving instrumental normativity, for which reason the arguments in the subsequent parts of the book should be quite convincing once one accepts the framing of the problems and the postulated ulterior ends (note that in (Schurz 1997, p.128) 'necessary' is understood differently, namely as *necessary by the laws of nature*). Regarding the latter (ii), $\Box(B \rightarrow A)$ means that from all available alternative means $B, C, D, \ldots$ to achieve A, B is optimal compared to $C, D, \ldots$. This way of considering normativity finds broad consensus in the practical domain (in approaches that suggest to maximise expected utilities, etc.) and seems to be a general feature of instrumental normativity. For this reason also this notion should not hinder one considering the arguments in our investigation as strong ones.

Second, although the schema shows that the tasks of postulating ulterior epistemic ends and epistemic engineering are distinct, it does not force one to consider them as completely independent. On the contrary, as we will see especially in part II of the book, sometimes the engineering part makes clear that there is no means which is generally accepted. Such a case (and *ought implies can*) suggests to revise one's ends then. So, not only that these tasks need not be independent, it even makes perfect sense to postulate some end ($\copyright A_1$), start the engineering and come up with some necessary means (B_1), revise one's end ($\copyright A_2$), start the engineering again (B_2), and so forth.

Finally, *epistemic engineering* is not only relevant for naturalised epistemology. In principle, also traditional approaches as, e.g., foundationalism, coherentism, and infinitism can be considered to perform epistemic engineering: Foundationalism, e.g., values non-circularity and finiteness ((EJ3) and (EJ4)) much higher than, e.g., coherentism does (values (EJ2) and (EJ4)). Similarly for infinitism which does not value (EJ4) much, but (EJ2) and (EJ3). Given these desiderata, they also engineer modifications of the strict notion of justification J, and they easily end up with different results as, e.g., externalist and internalist accounts of J, by valuing further desiderata differently. No problem with framing these positions this way, on the contrary, even better so for our making the case for epistemic engineering. Discussing naturalised epistemology served just as a well-known vehicle to characterise the instrumental approach to epistemic normativity and the engineer's perspective.

In the remainder of this part we introduce, so to say, some engineering tools. In parts II and III we then apply them in order to end up with statements about necessary and adequate means for the epistemic end of predictive success regarding a variety of problem settings. But for now, let us strike out on the *optimality route to justification*.

Chapter 2

The Setting

This chapter describes the setting in which all problems and solutions offered in this book are framed: the framework of online prediction games. *For this purpose, the elementary formal and mathematical tools which are subsequently employed are outlined. Then definitions that characterise online prediction games in detail are given. Since this investigation is after optimality claims, prominent measures of success and different notions of optimality are defined. Also a taxonomy of the epistemic methods under investigation is provided. At the end a branch of machine learning which is mainly concerned with defining optimal prediction methods is introduced. It is shown that the traditional problem of scepticism has a modern pendant in the machine learning literature. This is what allows one to exploit results of the machine learning literature for the aim of engineering a solution to the problem of epistemic justification in the subsequent parts of the book.*

From a normative standpoint, in science it is all about getting the truth. In empirical science this means that we need to get our explanations and predictions right. And for the other branches this means that we need to get our conceptual framework right. In our investigation we focus on the predictive part. There are some approaches in the philosophy of science which try to reduce the explanatory part to the predictive part. So, e.g., the debate about adequate criteria for explanations very often focuses on the ability of explanations to provide novel, unifying, etc. *predictions* (see, e.g., the overview in Schurz 2013, chpt.6). Although we subscribe to such approaches, we do not want to argue for such a reduction here. If one also subscribes to it, the better for a generalisation of our setting. If not, then it is the epistemic engineer's task to look out for further tools. Similarly for the conceptual framework part: In part II we outline applications of our setting to the framework of deductive and abductive reasoning. Getting along with these applications means that one also subscribes to a partial reduction of the conceptual framework part to the prediction part. Again, we consider such an approach promising, however, if it fails then it is again up to the epistemic engineer to find further tools.

As said before, we are after epistemic engineering regarding predictions. Very simplified speaking, the task of predicting consists in claiming something about the occurrence or non-occurrence of an event ahead of it in qualitative (e.g. 'will occur'/'will not occur'), comparative (e.g. 'is more likely to occur than') or quantitative (e.g. 'the probability of it occurring is') terms. E.g., if we represent whether an event occurred or not with values 0 and 1, then making a qualitative prediction amounts to providing one of these values before the event could have taken place or before its outcome was known. Predictive success comes with both values matching. Predictive failure with a mismatch between them. At least in the case of empirical science it is, so to speak, nature which fixes the first value and the scientist who fixes the latter. In this sense one might consider the task of prediction as a task of playing with (or against) nature. While this is the understanding in formal learning theory (K. T. Kelly 1996), in computational learning theory and machine learning different prediction algorithms are typically run against each other. In Schurz (2004, 2008b) this framework is described as a "prediction game". . Now, as we have outlined in the preceding chapter, we are more aiming at optimal than at true predictions. Speaking in figurative terms, nature is an opponent too strong to be beaten (we will argue for this in detail in chapter 5). Rather, the idea is to make optimal predictions in the sense that no other competitor outperforms one in terms of predictive success.

So, the setting of predictions we are working with are prediction games as studied in the theory of meta-induction. In the next section (2.1) we provide formal preliminaries of our study. Then, in section 2.2, we provide a formal characterisation of prediction games. Afterwards, we provide a formal characterisation of predictive success and optimality as well as a taxonomy of possible competitors in prediction games (section 2.3). Finally, we introduce the relevant theoretical background from the theory of meta-induction and machine learning and argue for our choice of this framework by help of methodological sceptical considerations as they are found in traditional epistemology (section 2.4).

2.1 A Basic Formal Toolbox

Although the formal apparatus of the theory of meta-induction is very advanced, we have restricted the investigation of this book to areas where in principle elementary tools can be employed which are basically covered by high-school mathematics and introductory classes on philosophical logic and probability theory. In particular, we will make use of methods of:

Elementary Logic. We use elementary logic, i.e. first order logic (and some propositional modal logic), for our explications and argumentation. In cases where formalism seems to us advantageous in order to make notions or claims generally more intelligible, we will make use of it—quite freely in both, the object- as well as the meta-language, as well as by mixing formal and natural language expressions. We use ordinary symbols for connectives ($\neg$: *not*, &: *and*, $\vee$: *or*, $\rightarrow$: *if-then*, $\leftrightarrow$: *if and only if / iff*), quantifiers ($\forall$: *all*, $\exists$: *some*, $\exists!$: *exactly one*), the identity relation ($=$: *identical*), and expressions for individuation such as the descriptive description operator ι ($(\iota x)(\varphi[x])$: that x which $\varphi[x]$, where $\varphi[x]$ means that x occurs in φ; for replacing all occurrences of x by y in the expression φ we will use later on $\varphi[x/y]$); we will use these symbols also in schematic expressions. Regarding modalities we use $\Box$ for *necessity* and *adequacy*, and $\mathcal{O}$ for *obligation* for the respective modality of any kind—the relevant kind should become clear in the respective context: logical, conceptual, metaphysical, epistemic, etc. We use $\vdash$ for deductive inferences and $\mid\!\sim$ for non-deductive/inductive inferences. In this section we use '$_{df}$' to mark definitions;

Naïve Set Theory. We operate on basis of naïve set theory with undefined $\in$, representing the *element relation* which is governed by *extensionality*:

$$\forall x, y\ (x = y \leftrightarrow \forall z\ (z \in x \leftrightarrow z \in y))$$

and *naïve comprehension*:

$$\exists Y \forall x (x \in Y \leftrightarrow \varphi[x])$$

Based on $\in$ we use as defined: *subset* and *proper subset* $\subset$,$\subseteq$: $X \subset Y$ iff$_{df}$ $X \neq Y$ & $\forall x (x \in X \rightarrow x \in Y)$ ($\subseteq$ allows for $X = Y$), *empty set* $\emptyset$: $\emptyset = X$ iff$_{df}$ $\forall y\ y \notin X$, *union*, *intersection* $\cup$,$\cap$: $X \cup / \cap Y = Z$ iff$_{df}$ $\forall z (z \in Z \leftrightarrow (z \in X \vee / \& z \in Y))$ (also the general forms $\bigcup, \bigcap$ defined on sets of sets), *difference* or *relative complement* $\setminus$: $X \setminus Y = Z$ iff$_{df}$ $\forall z (z \in Z \leftrightarrow (z \in X \& z \notin Y))$, *(absolute) complement* $^{\mathrm{C}}$: $X^{\mathrm{C}} = Z$ iff$_{df}$ $\forall z (z \in Z \leftrightarrow z \notin X)$, *power set* $\wp$: $\wp(X) = Y$ iff$_{df}$ $\forall Z (Z \in Y \leftrightarrow Z \subseteq X)$, *listing set brackets* $\{,\}$: $\{x_1, \ldots, x_n\} = X$ iff$_{df}$ $\forall x (x \in X \leftrightarrow (x = x_1 \vee \cdots \vee x = x_n))$, *set brackets with characteristic property* $\{x : \varphi[x]\} = X$ iff$_{df}$ $\forall y (y \in X \leftrightarrow \varphi[x/y])$, *tuple* $\langle x, y \rangle =_{df} \{\{x\}, \{x, y\}\}$, *n-tuple* $\langle x_1, \ldots, x_n \rangle =_{df} \langle \langle x_1, \ldots, x_{n-1} \rangle, x_n \rangle$ (recursively defined), *Cartesian product* $\times$: $X \times Y =_{df} \{z : \exists x \in X \exists y \in Y\ z = \langle x, y \rangle\}$, *n-ary Cartesian product* $X_1 \times \cdots \times X_n =_{df} \{z : \exists x_1 \in X_1 \ldots \exists x_n \in X_n\ z = \langle x_1, \ldots, x_n \rangle\}$, and finally $|X|$ for the *cardinality* of X defined via *equinumerousity* with the respective subset of $\mathbb{N}$.

High School Mathematics. In the following part we have collected the relevant parts of high school mathematics which are employed in this book; here are some meta-terminological stipulations:

- we use i, k, m, n as variables for integers (natural numbers $\mathbb{N}$);
- we use x, y, z as variables for real numbers ≥ 0, sometimes > 0 (most of the times we operate within $\in [0, 1]$ in this book);
- undefined arithmetical operations are $+$ (addition), $\cdot$ (multiplication); furthermore, basic are the set of natural numbers $\mathbb{N}$: $\{0, 1, 2, \dots\}$, and the set of (non-negative) real numbers $\mathbb{R}^{(+)}$;

Defined are *less* and *less or equal* $<, \leq$: $x < y$ iff$_{df}$ $\exists z \in \mathbb{R}^+ \; z \neq 0 \& y = x + z$ ($\leq$ allows for $= 0$); *greater than* and *greater or equal* $>, \geq$ as the inverse of $\leq, <$; *subtraction* $-$ as the inverse of $+$; *division* $/$ as the inverse of $\cdot$ (if $y > 0$, then $x/y = z$ iff$_{df}$ $x = y \cdot z$); the *absolute function* $|x|$ defined as: $|x| = x$ if $x \geq 0$, and $|x| = 0 - x$ if $x < 0$; *proportionality* $x \propto y$ iff$_{df}$ $\exists z \; x = y \cdot z$; the *closed interval* $[x, y] =_{df} \{z : x \leq z \leq y\}$; furthermore, maximum and minimum functions for selecting the *maximum/minimum according to domain*: max / min$(x_1, \dots, x_n) =_{df} (\iota y)((y = x_1 \vee \dots \vee y = x_n) \& y \geq / \leq x_1 \& \dots \& y \geq / \leq x_n)$; and likewise for selecting the *maximum/minimum according to image*: $\arg\max_{x \in X} / \arg\min_{x \in X} f(x) =_{df} \{y : y \in X \& \forall z \in X \; f(y) \geq / \leq f(z)\}$; defined is also the *averaging function* (overline): $\overline{f(x)}_{x \in X} =_{df} \frac{\sum_{x \in X} f(x)}{|X|}$; the rounding functions $\lfloor, \rfloor$ (*round down*) $\lceil, \rceil$ (*round up*), and $[]$ (*round to the next integer (half up)*): $\lfloor x \rfloor =_{df} (\iota n \in \mathbb{N})(n \leq x \& \forall m \in \mathbb{N}(m \leq x \rightarrow m \leq n))$, analogously for $\lceil x \rceil$, and $[x] =_{df} \lfloor x \rfloor$, if $x - \lfloor x \rfloor < 0.5$ and $[x] =_{df} \lceil x \rceil$ otherwise;

Furthermore, defined are the following notions:

Sequence. A sequence $\langle x_n \rangle$ is an ordered collection of objects/numbers in which elements can repeatedly occur. E.g. $\langle x_n \rangle = \langle 0, 2, 4, 6, \dots \rangle$ is the sequence of even natural numbers ordered according to $<$. We can define such a sequence by providing a characteristic property of its elements as, e.g.: $\langle x_n \rangle : 2 \cdot n$ ($n \in \mathbb{N}$), or recursively: $x_{n+1} = x_n + 2$ (with $x_0 = 0$ and $n \in \mathbb{N}$). Now, there is one property of sequences we are mainly interested in this book, namely *convergence*. A sequence converges, if *almost* all elements of the sequence are *arbitrarily* close to some value. If there is such a value, one also says that the sequence *converges to* that value. In mathematics *arbitrarily close* means that for *any* real number > 0 the elements of the sequence get close to the value; and *almost all* means that there is only a finite number of exceptional cases. Formally, we can express this, e.g., with the notion of a *limit* of a sequence:

If $\exists x \forall \varepsilon > 0 \exists m \forall n \geq m : |x_n - x| < \varepsilon$, then:

$$\lim_{n \to \infty} (\langle x_n \rangle) = x \;\text{ iff}_{df}\; \forall \varepsilon > 0 \exists m \forall n \geq m : |x_n - x| < \varepsilon$$

Often, also the angle brackets are suppressed in the notation; if so, one writes '$\lim_{n \to \infty} (x_n)$'. Clearly, $\langle x_n \rangle : 2 \cdot n$ has no limit, because for any y of $\langle x_n \rangle$

there are infinitely many z of $\langle x_n \rangle$ such that for some $\varepsilon > 0$: $|y - z| > \varepsilon$. But also, e.g., $\langle x_n \rangle : -1^n$ which is $\langle 1, -1, 1, -1, \dots \rangle$ has no limit, because what is arbitrarily close to 1 is not arbitrarily close to -1 and vice versa, and for both values there are infinitely many elements in the sequence (note, however, that both sequences are *bounded from below*/have a *lower bound*, namely 0 and -1 respectively, but, whereas the second sequence is also *bounded from above*/has as *upper bound* 1, the sequence of even numbers has no upper bound; in such a case we will also write from time to time $\lim\limits_{n \to \infty} x_n = \infty$). Also clearly, $\langle x_n \rangle : \frac{1}{n}$ has a limit, namely 0: Consider any $\varepsilon > 0$. Then there is an $m \in \mathbb{N}$ such that $\frac{1}{m} < \varepsilon$ (by the *Archimedean property* of the real numbers $\mathbb{R}$). Hence, also $|\frac{1}{m} - 0| < \varepsilon$. Now, for all $n \geq m$: $\frac{1}{n} \leq \frac{1}{m}$, hence $|\frac{1}{m} - 0| \leq |\frac{1}{m} - 0|$, hence $|\frac{1}{n} - 0| < \varepsilon$. So 0 is *the* limit of $\langle x_n \rangle : \frac{1}{n}$. Any such sequence $\langle x_n \rangle$ with $\lim\limits_{n \to \infty} (x_n) = 0$ is called a *null sequence*.

Now, already from the definition above and the latter definitions the following important rules for arithmetic operations with the limit follow (note that the arithmetical operations on the sequences are defined via applying the operations for all elements of the sequence as, e.g., $\langle x_n + y_n \rangle = \langle x_1 + y_1, x_2 + y_2, \dots \rangle$, $\langle y \cdot x_n \rangle = \langle y \cdot x_1, y \cdot x_2, \dots \rangle$, etc.): Let $\langle x_n \rangle$ and $\langle y_n \rangle$ have a limit, then:

$\lim\limits_{n \to \infty} (x_n + y_n) = \lim\limits_{n \to \infty} (x_n) + \lim\limits_{n \to \infty} (y_n)$ (also for subtraction $-$)
$\lim\limits_{n \to \infty} (y \cdot x_n) = y \cdot \lim\limits_{n \to \infty} (x_n)$
$\lim\limits_{n \to \infty} (x_n \cdot y_n) = \lim\limits_{n \to \infty} (x_n) \cdot \lim\limits_{n \to \infty} (y_n)$
$\lim\limits_{n \to \infty} (\frac{x_n}{y_n}) = \frac{\lim\limits_{n \to \infty} (x_n)}{\lim\limits_{n \to \infty} (y_n)}$ (if 0 is not in $\langle y_n \rangle$ and $\langle y_n \rangle$ is no *null sequence*)
$\lim\limits_{n \to \infty} (x_n^y) = \left(\lim\limits_{n \to \infty} (x_n) \right)^y$
If $\exists m \forall n x_n \leq y_m$, then $\lim\limits_{n \to \infty} (x_n) \leq \lim\limits_{n \to \infty} (y_n)$

Note that $\langle x_n \rangle$, $\langle y_n \rangle$ might have no limit, whereas $\langle x_n \odot y_n \rangle$ might have one (e.g.: If $\langle y_n \rangle = \langle x_n \rangle : n$; then $\langle x_n \rangle$ as well as $\langle y_n \rangle$ have no limit, although, e.g., the limit of $\langle x_n - y_n \rangle$ is 0).

Summation. Next to the limit, we will often make use of summation. It is defined as follows:

- Basis: $\sum\limits_{i=m}^{n} x_i =_{df} 0$ if $m > n$
- Recursion: $\sum\limits_{i=m}^{n} x_i =_{df} x_n + \sum\limits_{i=m}^{n-1} x_i$ if $m \leq n$

- Infinity: $\sum\limits_{i=1}^{\infty} x_i =_{df} \lim\limits_{n\to\infty} \sum\limits_{i=1}^{n} x_i$
  (The definition for the infinite case is conditional on the existence of the limit.)

Note that strictly speaking the summation operation is defined on sequences (we suppressed the angle brackets). The sum of a sequence is also called a *series*. Important arithmetical operations on series are:

$\sum\limits_i (x_i + y_i) = \sum\limits_i (x_i) + \sum\limits_i (y_i)$	$\sum\limits_i (y \cdot x_i) = y \cdot \sum\limits_i (x_i)$
$\sum\limits_{i=1}^{n} x = n \cdot x$	$\sum\limits_{i=m}^{n} x = (1 + n - m) \cdot x$
Special case: $\sum\limits_{i=1}^{n} i = \frac{n\cdot(n+1)}{2}$	

A sequence/series is *arithmetic*, if the difference between neighbouring elements remain constant: $\exists y : x_{n+1} = x_n \pm y$. E.g. $\langle 0, 2, 4, 6, \dots \rangle$ is an arithmetic sequence, $\sum\limits_{i=0}^{n} 2 \cdot i$ is an arithmetic series. A sequence/series is *geometric*, if the ratio between neighbouring elements remains constant: $\exists y : x_{n+1} = y \cdot x_n$. E.g. $\langle 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots \rangle$ is a geometric sequence, $\sum\limits_{i=0}^{n} \frac{1}{2}^i$ is a geometric series.

Product. Analogously to summation, we can define also the general product:

- Basis: $\prod\limits_{i=m}^{n} x_i =_{df} 1$ if $m > n$
- Recursion: $\prod\limits_{i=m}^{n} x_i =_{df} x_n \cdot \prod\limits_{i=m}^{n-1} x_i$ if $m \leq n$
- Infinity: $\prod\limits_{i=1}^{\infty} x_i =_{df} \lim\limits_{n\to\infty} \prod\limits_{i=1}^{n} x_i$
  (The definition for the infinite case is conditional on the existence of the limit.)

E.g. it holds: $\prod\limits_{i=1}^{n} x = x^n$; as special case we can define the *faculty* of an integer n:

$$n! =_{df} \prod_{i=1}^{n} i \quad (= 1 \cdot 2 \cdot \dots \cdot n)$$

Exponentiation. Exponentiation is defined in several steps for different domains: integer exponentiation is defined as: $x^n =_{df} \prod\limits_{i=1}^{n} x$; here x is also called the '*basis*', and n the '*exponent*'; exponentiation with rational numbers is defined as: $x^{\frac{m}{n}} = y$ iff$_{df}$ $x^m = \prod\limits_{i=1}^{n} y$; the general case of exponentiation with a real number is defined as: $x^y = z$ iff$_{df}$

- $\exists\langle z_n\rangle$: $\langle z_n\rangle$ is a sequence of rational numbers $(\frac{n_1}{m_1}, \frac{n_2}{m_2}, \ldots)$ with $\lim_{n\to\infty}(z_n) = y$, and:
- $z = \lim_{n\to\infty} x^{\langle z_n\rangle}$

Important calculation rules for exponentiation are:

$x^0 = 1$	$x^{-y} = \frac{1}{x^y}$	$x^{\frac{1}{y}} = \sqrt[y]{x}$	$x^{y+z} = x^y \cdot x^z$
$x^{y-z} = \frac{x^y}{x^z}$ $(x^{-y} = \frac{1}{x^y})$	$\left(\frac{x}{y}\right)^z = \frac{x^z}{y^z}$	$(x^y)^z = x^{y\cdot z}$	$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$

e is *Euler's number*, where $e = e^1 \approx 2.71828$. Sometimes we will write '$\exp(x)$' for 'e^x'.

Root. One inverse operation of exponentiation concerns an inversion function with respect to the exponent, which is the root: If $x \leq 0$, then:

$$\sqrt[n]{x} = y \ \text{iff}_{df} \ y^n = x$$

Important calculation rules for the root are (if the degree n of the root is not relevant, it is omitted):

$\sqrt{x\cdot y} = \sqrt{x}\cdot\sqrt{y}$	$\sqrt{\frac{x}{y}} = \frac{\sqrt{x}}{\sqrt{y}}$	$\sqrt{x^n} = (\sqrt{x})^n$
$\sqrt[n]{x}\cdot\sqrt[m]{x} = \sqrt[n\cdot m]{x^{n+m}}$	$\sqrt[m]{\sqrt[n]{x}} = \sqrt[n\cdot m]{x}$	$\sqrt[n]{x} = x^{\frac{1}{n}}$

Logarithm. The other inverse operation of exponentiation concerns an inversion function with respect to the basis, which is the logarithm:

$$\log_z(x) = y \ \text{iff}_{df} \ z^y = x$$

Important calculation rules for the logarithm are (if the basis z of the logarithm is not relevant, it is omitted):

$\ln(x) \ =_{df} \ \log_e(x)$	$\text{lb}(x) \ =_{df} \ \log_2(x)$
$\lg(x) \ =_{df} \ \log_{10}(x)$	$\log(1) = 0$
$\log(x\cdot y) = \log(x) + \log(y)$	$\log\left(\prod_{i=1}^{n} x_i\right) = \sum_{i=1}^{n}(\log(x_i))$
$\log\left(\frac{x}{y}\right) = \log(x) - \log(y)$	$\log(x^y) = y\cdot\log(x)$
$\log(\sqrt[y]{x}) = \frac{1}{y}\log(x)$	$\log_y(x) = \frac{\log_z(x)}{\log_z(y)}$

Important (In-)Equalities. The following inequalities/equalities are important for some proofs of upper bounds of learning algorithms (chapter 3):

- $e^{-x} \leq 1 - x + \frac{x^2}{2}$ (valid for all $x \geq 0$)
- $e^{-x} \geq 1 - x$ (valid for any x)
- Geometric sum: $\forall x \neq 1$:

$$\sum_{i=0}^{n} x^i = \frac{1 - x^{n+1}}{1 - x}$$

Combinatorics. From time to time we will apply general results of the machine learning literature to particular cases and illustrate them by combinatorial considerations. Most important is the case of selecting items from a collection where the order does not matter, i.e. the case of *combination*. For a selection of k out of n different elements where the order does not matter, i.e. for a combination of k out of n elements without repetition, there is $\binom{n}{k}$ possibilities:

$$\binom{n}{k} =_{df} \frac{n!}{k! \cdot (n-k)!}$$

$\binom{n}{k}$ reads as *n choose k*, and is also called a *binomial coefficient*, because it is also the coefficient of the $x^k y^{n-k}$-term in the so-called *polynomial expansion* of the binomial power $(x+y)^n$. E.g.: $(x+y)^2 = \binom{2}{0}x^2y^0 + \binom{2}{1}x^1y^1 + \binom{2}{2}x^0y^2 = x^2 + 2xy + y^2$. Important calculation rules are:

If $k > n$, then $\binom{n}{k} =_{df} 0$	$\binom{n}{0} = 1$	$\binom{n}{n} = 1$
$\binom{n}{1} = n$	$\binom{n}{n-1} = n$	$\binom{n}{n-k} = \binom{n}{k}$
$\binom{n}{k-1} = \binom{n}{k} \cdot \frac{k}{n-k+1}$	$\binom{n+1}{k} = \binom{n}{k} \cdot \frac{n+1}{n+1-k}$	$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$

Probability Theory. We start with a set of elementary events or event types $Y = \{Y_1, \ldots, Y_n\}$. Next we choose an algebra $\mathcal{A}$ over Y, which is a subset of $\wp(Y)$ that contains Y and is closed under ^{C} (complement), $\cup$ (union), and $\cap$ (intersection). I.e.:

($\mathcal{A}$1) $Y \in \mathcal{A}$

($\mathcal{A}$2) $\mathcal{A} \subseteq \wp(Y)$

($\mathcal{A}$3) $\forall X(X \in \mathcal{A} \rightarrow X^C \in \mathcal{A})$

($\mathcal{A}$4) $\forall X \forall Z((X \in \mathcal{A} \& Z \in \mathcal{A}) \rightarrow X \cup Y \in \mathcal{A})$

($\mathcal{A}$5) $\forall X \forall Z((X \in \mathcal{A} \& Z \in \mathcal{A}) \rightarrow X \cap Y \in \mathcal{A})$

The most fine-grained algebra over Y is $\wp(Y)$, the most coarse-grained one is Y itself. For illustrative purpose, consider the case of throwing an ordinary dice. The so-called *possibility space* of event outcomes is $\{1,2,3,4,5,6\}$; we can characterise a set of event types $Y = \{Y_1, \ldots, Y_6\}$ via Y_i: *The dice lands on i*. If we choose as algebra $\mathcal{A}$ over Y: $\mathcal{A} = Y$, then we can distinguish only between the safe event Y that the dice lands on one out of $\{1,2,3,4,5,6\}$ and the excluded event $\emptyset$ that it does not land on one out of $\{1,2,3,4,5,6\}$. If we assume, e.g., $\mathcal{A} = \{\{Y_1\}, \{Y_2, \ldots, Y_6\}, Y\}$, then we can distinguish also between landing on 1 or not. And if we assume $\mathcal{A} = \wp(Y)$, then we can distinguish all combinatorial possible cases. The probabilistic considerations of this book are always about the most fine-grained algebra over the set of *elementary events* or *event types*.

In philosophy, probability theory is often applied on an algebra not over events, but propositions or statements. Such a sentential algebra is defined over a set of elementary propositions $\{p_1, p_2, \ldots, \}$ and is closed under $\neg$, &, and $\vee$ (a *Boolean algebra*). Since probability theory is applied equivalently to an algebra of events as to an algebra of propositions, we will also switch between these two types when applying probability theory (e.g., the former type is more relevant when we speak about random variables).

On top of such an algebra, we can define a probability distribution *Pr* as any function satisfying the so-called *Kolmogorov*-axioms, first put forward by Andrey N. Kolmogorov in 1933: Let $A, B \in \mathcal{A}$, then:

- Unconditional *Pr*:

  (Pr1) *Non-negativity*: $Pr(A) \geq 0$

  (Pr2) *Normalisation*: $Pr(\top) = 1$ (for any tautology or safe event $\top$)

  (Pr3) *(Finite) Additivity*: $Pr(A \vee B) = Pr(A) + Pr(B)$, if A and B are mutually exclusive or logically contrary

- Conditional *Pr*:

  (Pr4) Given $Pr(A) > 0$, then:

$$Pr(B|A) =_{df} \frac{Pr(B\&A)}{Pr(A)}$$

The conditional probability $Pr(B|A)$ is also described as the probability of B in the light of A which is thought to express the probability of B once one got to know or is certain about A. So, e.g., if A logically implies B: $A \vdash B$, then, given A or in the light of A, B is certain, i.e: $Pr(B|A) = 1$. Given a probability distribution *Pr*, we can "check" for probabilistic dependencies and independencies between events or propositions simply by considering their probabilistic “impact”: If B turns out to be less probable in the light of A than unconditioned on A, i.e. if $Pr(B|A) < Pr(B)$, then we say that A and B are *negatively correlated*; note that if $Pr(B|A) < Pr(B)$, then also $Pr(A|B) < Pr(A)$. If B turns out to be more probable conditional on A, i.e. if $Pr(B|A) > Pr(B)$, then we say that A and B are *positively correlated* (again, $Pr(B|A) > Pr(B)$ implies $Pr(A|B) > Pr(A)$). If A has no positive or negative probabilistic “impact” on B, i.e. if $Pr(B|A) = Pr(B)$, then we say that A and B are *probabilistically independent* (also here $Pr(B|A) = Pr(B)$ implies $Pr(A|B) = Pr(A)$); equivalently, probabilistic independence can be characterised via: $Pr(A\&B) = Pr(A) \cdot Pr(B)$.

Important calculation rules are:

Negation theorem:	$Pr(\neg A) = 1 - Pr(A)$
Consequence theorem:	If $B \vdash A$, then $Pr(B) \geq Pr(A)$
Equivalence theorem:	If $A \vdash\dashv B$, then $Pr(A) = Pr(B)$
Chain rule:	$Pr(A_1\& \cdots \&A_n) = \prod_{i=1}^{n} Pr(A_i\|A_1\& \cdots \&A_{i-1})$
Law of total probability:	If $\vdash (B_1 \dot\vee \cdots \dot\vee B_n)$, then $Pr(A) = \sum_{i=1}^{n} Pr(A\&B_i) = \sum_{i=1}^{n} Pr(A\|B_i) \cdot Pr(B_i)$ (note that $\dot\vee$ is *xor*, the exclusive disjunction, i.e. the B_is are pairwise exclusive and jointly exhaustive; a special case is: $Pr(A) = Pr(A\&B) + Pr(A\&\neg B)$)
Bayes’ theorem:	$Pr(A\|B) = Pr(B\|A) \cdot \frac{Pr(A)}{Pr(B)}$

When speaking of probabilities of events, we make use of *random variables* X, Z, also with sub-indices. A random variable predicates over individuals properties from a set of mutually exclusive and jointly exhaustive properties. E.g. if $\{F(\cdots), \neg F(\cdots)\}$ is the set of properties under consideration, then X might predicate $F(\cdots)$ to event e, i.e. $X(e)$ is equivalent to $F(e)$. Quite often the properties under consideration are trivial identifications—as, e.g., in the case of rolling a dice where X predicates about the rolling event e one of $\{1, 2, 3, 4, 5, 6\}$, more specifically, one of the set of properties $\{\cdots = 1, \ldots, \cdots = 6\}$. In such a case, if it is clear from the context which event is referred to, we will write ‘$X = i$’ ($1 \leq i \leq 6$) or, more generally, ‘$X \in \mathbb{R}$’ instead of ‘$\underbrace{X(e)}_{i=} e$’.

We will make use of discrete random variables only in this book, which means that the set of mutually exclusive and jointly exhaustive properties under consideration is finite; if $X \in \mathcal{V} \subset \mathbb{R}$, where $\mathcal{V}$ is also called the *value space of* X and is finite, then we can define the so-called *expected value* of X relative to $\mathcal{V}$ and a probability function Pr as:

$$\mathbb{E}[X] =_{df} \sum_{v \in \mathcal{V}} Pr(X = v) \cdot v$$

Intuitively, the expected value of X is the long-run average value of repetitions of events covered by X. E.g., if X is about rolling a fair dice with $\mathcal{V} = \{1,2,3,4,5,6\}$ and $Pr(X = 1) = \cdots = Pr(X = 6) = \frac{1}{6}$, then $\mathbb{E}[X] = 3.5$ which is also the average of $\{1,2,3,4,5,6\}$. If, e.g., the dice has a strong bias towards 6: $Pr(X = 6) = \frac{3}{6}$ and $Pr(X = 1) = \cdots = Pr(X = 5) = \frac{3}{5 \cdot 6}$, then $\mathbb{E}[X] = 4.5$.

Considering more than one random variable, we might be interested in their interactions. Many important theorems concern the case where random variables are attached to the same (*identical*) probability distribution without there being a probabilistic influence of one to the other (*independent*). E.g., when rolling one and the same dice twice, it is assumed that the possible outcomes of the second roll are equally probable to the possible outcomes of the first roll. It is furthermore assumed that the probability of a possible outcome of the second roll is not influenced by a (possible) outcome of the first roll and vice versa. So, if X_1 covers the first roll and X_2 the second one with $\mathcal{V}_1 = \mathcal{V}_2 = \{1, \ldots, 6\}$, we usually assume that $Pr(X_1 = i) = Pr(X_2 = i)$, i.e. X_1 and X_2 are identically probabilistically distributed, and $Pr(X_1 = i | X_2 = j) = Pr(X_1 = i)$, i.e. X_1 and X_2 are independently distributed. If this is the case, one also speaks of *independent and identically distributed* (i.i.d.) random variables. An important statistical theorem we will employ in this book concerns an infinite sequence of such random events $X_1, X_2, \ldots$ which are pairwise i.i.d.: The so-called *weak law of large numbers* states that the arithmetic mean of the values almost surely (i.e. with probability one) converges (i.e. gets arbitrarily close) to the expected value as the number of repetitions of such an event approaches infinity ($v_1, \ldots, v_n$ are the outcomes/values *de facto* assigned by $X_1, \ldots, X_n$ to the events):

$$\forall \varepsilon > 0 : \lim_{n \to \infty} Pr\left(\left| \frac{v_1 + \cdots + v_n}{n} - \mathbb{E}[X_1] \right| < \varepsilon \right) = 1$$

Bayesian Networks. Finally, we will also sometimes illustrate and apply the general results of meta-induction in cases best described by help of so-called *Bayesian networks*. Such networks allow for graphically representing the paths over which probabilistic information spreads between random variables. They consist of a set **V** of random variables $X_1, \ldots, X_n$, a set **E** of directed edges ($\longrightarrow$) connecting some of these variables, and a probability function Pr over $\mathbf{V}$. A triple $\langle \mathbf{V}, \mathbf{E}, Pr \rangle$ is a Bayesian network iff it conforms to the so-called *Markov factorisation* (Pearl 2000, p.16)

$$Pr(X_1 = v_1 \& \cdots \& X_n = v_n) = \prod_{i=1}^{n} Pr(X_i = v_i | \mathbf{Par}(X_i)), \tag{2.1}$$

where $\mathbf{Par}(X_i)$ is the set of X_i's "parents" in the Bayesian network's graph $\mathbf{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, i.e., the set of all $X_j \in \mathbf{V}$ for which $X_j \longrightarrow X_i$ holds. Whenever the probability distribution Pr of a triple $\langle \mathbf{V}, \mathbf{E}, Pr \rangle$ factors according to Markov factorisation (equation (2.1)), then one can read off certain independencies in Pr from the graph $\mathbf{G} = \langle \mathbf{V}, \mathbf{E} \rangle$. In particular, every $X_i \in \mathbf{V}$ has to be independent of every X_j that is not connected to X_i via a path $X_i \longrightarrow \cdots \longrightarrow X_j$ conditional on $\mathbf{Par}(X_i)$. Typically applied are Bayesian networks, e.g., when causally interpreted, i.e. when the arrows ($\longrightarrow$) of a Bayesian network's graph stand for direct cause-effect relationships. In this book, however, we will employ Bayesian networks only with respect to the aim of simpler reading off and representing probabilistic information. Figure 2.1 illustrates a Bayesian network.

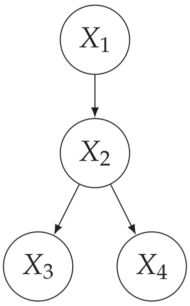

Figure 2.1: Example of a Bayesian network: It encodes the probabilistic information that X_3 is probabilistically independent of X_4 as well as X_1—both times conditional on X_2; analogously for X_4. ($\mathbf{Par}(X_3) = \mathbf{Par}(X_4) = \{X_2\}$, $\mathbf{Par}(X_2) = \{X_1\}$, $\mathbf{Par}(X_1) = \emptyset$).

2.2 Prediction Games

Let us start with some kind of *atomism* by assuming that "the world divides into facts" (Wittgenstein 1961, prop.1.2, p.7). Now, actually we want to take an *epistemic* stance of atomism, which is to say that we assume the world to be *distinguishable* in *events* Y. Regarding prediction tasks, we are interested in the prediction of a series of events of one and the same type, so we want to distinguish *event types* Y^s: $Y^1, Y^2, \ldots$. Think, e.g., on types of stocks as, e.g., *Apple stock*, *Google stock*, etc., or different types of weather as, e.g., *rainy*, *sunny*, etc. But we are also interested in temporal relations. For this reason we also distinguish events Y according to the time of their possible occurrence into $Y_1, Y_2, \ldots$, the (possible) event tokens. Think, e.g., on a particular Apple stock or a particular rainy day, etc. How event types and points in time are distinguished exactly, is left open. There is no need of making any specific restrictions to natural kinds or some other forms of filtering out specific types. The only thing that matters is that there are possibly infinitely many instances of event types in time. In order to keep technicalities simple, we also assume that for each event the numbers of instances in time are only countably many ($\mathbb{N}$), however, theoretically nothing hinges on this assumption and one might transfer our results to the case of dense points in time ($\mathbb{R}$). By combining both ways of distinction as Y_t^s, we describe the world as a matrix as depicted in figure 2.2.

Y_1^1	Y_2^1	Y_3^1	Y_4^1	Y_5^1	$\ldots$
Y_1^2	Y_2^2	Y_3^2	Y_4^2	Y_5^2	$\ldots$
Y_1^3	Y_2^3	Y_3^3	Y_4^3	Y_5^3	$\ldots$
Y_1^4	Y_2^4	Y_3^4	Y_4^4	Y_5^4	$\ldots$
Y_1^5	Y_2^5	Y_3^5	Y_4^5	Y_5^5	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Figure 2.2: *The* world as a *mosaic* or the totality of events: We use super-indices to refer to event types and sub-indices to refer to points in time. By mixing both we refer to (possible) event tokens or simply '(possible) events'.

Now, we assume that one can always describe the outcomes of such an event in quantitative terms. In case one just wants to represent whether an event took place or not, one can simply use binary values 0, 1: Y_2^4, e.g., might be mapped to 1 if an event of type Y^4 (e.g. *rainy*) took place at point in time 2; otherwise it is mapped to 0. In case one wants to transform a more general qualitative description of the output into a quantitative one, one can do so by *retyping* the events and binarising each qualitative value separately. If, e.g., one wants to speak of possible outcomes *sunny*, *rainy* and *windy* (having a more general type *weather* in mind), then one can distinguish this event type into type Y^3 (*sunny*), type Y^4 (*rainy*) and type Y^5 (*windy*) and proceed as above, i.e. describe the outcome via $\in \{0, 1\}$. However, the outcome need not be only binary, but can take on any real value one wants—as is the case, e.g., with stock values. We only assume that generally such quantified event outcomes can be normalised to $[0, 1]$. This assumption is unproblematic as long as we assume that the quantities we are speaking about are somehow *bounded*. To sum up, we assume that event outcomes are quantitatively described and bounded. The formal details are provided below in definition 2.4.

Taking this epistemic point of view, it is easy to describe the task of making predictions about the world in game-theoretical terms: A prediction task consists in an individual's i making a claim f at some point in time t about the quantified outcome y of an event Y occurring at some later point in time $> t$. If we take y_t^s to be a value *provided by* the world or nature and $f_{t,i}^s$ to be a value provided by an individual i, then a prediction task is a game between nature and individual i. In a round of such a game i wins if the prediction succeeds, i.e. if $f_{t,i}^s = y_t^s$, and i looses against nature if the prediction fails, i.e. if $f_{t,i}^s \neq y_t^s$. One can also provide a quantitative measure for winning and loosing per round by introducing a loss function ℓ. This function is supposed to measure somehow something which somehow might be interpreted as something like a distance between i's prediction and nature's choice: $\ell(f_{t,i}^s, y_t^s)$. Note that we opted for such an awkward expression, because we do not really assume that ℓ is a distance measure (i.e. non-negative, symmetric, subadditive, and indiscernible regarding identicals), although this would be a quite natural constraint. It is also common to interpret ℓ as a *loss function*, measuring the loss a player receives at a round for her prediction in a prediction game (for details see below). For technical convenience we assume that ℓ is indiscernible regarding identicals in the following way:

Axiom 2.1 (Loss)**.**

$$\ell(x, x) = 0 \quad \forall x \in [0, 1]$$

Furthermore, we assume that ℓ operates also within $[0, 1]$, so, we assume that ℓ is bounded:

Axiom 2.2 (Loss)**.**

$$\ell(x, z) \in [0, 1] \quad \forall x, z \in [0, 1]$$

By taking the inverse of the loss, one gets a measure for success:

Definition 2.3 (Simple Success)**.**

$$s(x, z) = 1 - \ell(x, z) \quad \forall x, z \in [0, 1]$$

s also operates within $[0, 1]$; we will discuss several such success measures in the subsequent section.

We use such a measure of predictive success for evaluating predictions and their underlying methods. As we will see soon, non-trivial evaluation cannot be performed in absolute terms. In order to be non-trivial, a general evaluation procedure always needs some standards for comparison. In the case of making a prediction, this is usually done by considering a concatenation of predictions or just repeated predictions. We aim to reach a comparative evaluation, for which reason we will consider concatenations of predictions of several individuals. The latter is provided in the framework of *prediction games* which we introduce now by generalising the notion provided in (Schurz 2008b).

A prediction game consists of the following ingredients:

Definition 2.4 (Events, Predictions, and Truth)**.**

- Y_t^s: $Y_1^1, Y_2^1, \ldots; Y_1^2, Y_2^2, \ldots$ are infinite series of events.
- $\mathcal{Y} = \langle\langle y_1^1, y_2^1, \ldots\rangle, \langle y_1^2, y_2^2, \ldots\rangle, \ldots\rangle$ are quantified representations (within the interval $[0,1]$) of the true (or actual) outcomes (or values) of the events (event variables) to be predicted: $y_t^s \in [0,1]$.
- F_i: $F_1, \ldots, F_n$ are the prediction or forecasting methods of $n \in \mathbb{N}$ predictors or forecasters.
- $\mathcal{F} = \langle\langle\langle f_{i,1}^1, f_{i,2}^1, \ldots\rangle, \langle f_{i,1}^2, f_{i,2}^2, \ldots\rangle, \ldots\rangle : 1 \leq i \leq n\rangle$ are the predictions or forecasts of the single events within the interval $[0,1]$ of the predictors or forecasters $1 \leq i \leq n$: $f_{i,t}^s \in [0,1]$

More precisely, we define a prediction game by the following 4-tuple:

Definition 2.5 (Prediction Game)**.** *G* is a prediction game (with the true values $\mathcal{Y}$ and the predicted values $\mathcal{F}$) about events of type(s) $I \subseteq \mathbb{N}$ iff

$$\begin{aligned} G = \langle\ & \{\langle s, t, Y_t^s\rangle : t \in \mathbb{N} \;\&\; s \in I\}, \\ & \{\langle s, t, y_t^s\rangle : t \in \mathbb{N} \;\&\; s \in I\}, \\ & \{F_i : 1 \leq i \leq n\}, \\ & \{\langle s, i, t, f_{i,t}^s\rangle : 1 \leq i \leq n \;\&\; t \in \mathbb{N} \;\&\; s \in I\}\ \rangle \end{aligned}$$

I is a set of indices of the event types in question ($I \subseteq \mathbb{N}$). E.g., a prediction game with $I = \mathbb{N}$ amounts to a task of predicting everything (assuming that the set of all properties is countably infinite, as is done, e.g., according to our approach presented in figure 2.2); $I = \{3,4,5\}$ filters out a prediction game on weather; one might put forward probabilistic constraints for connecting the predicted values $f_{i,t}^3, f_{i,t}^4, f_{i,t}^5$ as well as for the outcomes y_t^3, y_t^4, y_t^5 such that they are non-negative and sum up to 1 and for the *y*s one typically assumes that they are $\in \{0, 1\}$ (for all $i, t \in \mathbb{N}$); given these constraints, one can interpret such a prediction game as a probabilistic one. On the other hand, setting $I = \{3\}$ filters out a simple prediction game on all events of type Y^3 (whether it is sunny or not or to which degree it is sunny etc.). If I is a singleton and the specific event type is irrelevant, then we will just omit super-indices. We will speak then also about a 'prediction game' simpliciter.

Note that our notion of a prediction game generalises that of Schurz (2008b) insofar we include several sequences of different types of events. The same holds also for the structure of forecasts and forecasters: In our setting they are about event variables, whereas in the original setting of Schurz they are about single events. However, in most parts of the book our setting coincides with that of Schurz (2008b) as we speak mainly about prediction games with $|I| = 1$.

Relevant for the evaluation of predictions within a prediction game are especially the values of y and f_i (see figure 2.3): The closer $f^s_{i,t}$ is to y^s_t, the better the prediction of i. And the closer the $f^s_{i,t}$'s are to the respective y^s_t's, the better i is a predictor in general. Although reference to events themselves as well as reference to the forecasting methods is not directly relevant for evaluation, we have included them into our definition of prediction games in order to be able to make more distinctions of such games relevant for our later applications. Thereby we consider a prediction method F_i to be a function that maps some input, including the event in question, into the predicted output value:

y_1	y_2	y_3	y_4	y_5	$\dots$
$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	$f_{1,5}$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$f_{n,1}$	$f_{n,2}$	$f_{n,3}$	$f_{n,4}$	$f_{n,5}$	$\dots$

Figure 2.3: Prediction game with event outcomes y and predictions $f_i, \dots, f_n$ of n predictors

$$F_i : \text{Input}(s, t) \longrightarrow f^s_{i,t} \in [0, 1]$$

Depending on the exact specification of Input, one can differentiate different kinds of prediction methods. If, e.g., Input includes, besides Y^s_t, all or some outcomes of past events of type Y^s, but none of future events ($\{\langle s, u, y^s_u\rangle : u < t\}$), then F_i is an *object level* method (e.g., an *inductive* method). If it includes next to Y^s_t outcomes of past events of other types ($Y^{\neq s}$), then F_i might be an *abductive* method or a method employing *analogies* and *simulations*. However, formally similarly it might be also some *oracle* or *witchcraft* method, employing empirical data on other event types like extispicy in order to make a prediction about a future event like the outcome of a war. If F_i includes also the outcome of the future event (y_t^s), then F_i might be interpreted as a *clairvoyant* method, since its application presupposes access to the future. Since the enlightenment, but also already before, the later kind of methods were tackled especially by science. For this reason one might speak of *para-scientific methods* here. If no event outcome is included at all, then F_i is an *a priori* method. One such method could be, e.g., a method that produces its predictions *randomly*. Another possibility is, e.g., a purely "rationalistic method" that *deduces* its predictions only from *first principles* that are not empirical. There are further possibilities for the choice of Input and at the end of this chapter we will provide a taxonomy of all prediction methods relevant for our endeavour (see section 2.3). However, for our current purpose we do not need to make any restrictions in this respect. Prediction games can consist of any prediction methods one can think of. The only thing that matters is that such a method provides predictions of the form $f_{i,t}^s$, because the latter can be used for evaluation by measuring their deviation from the true outcome, which is also a measure for success. Now, before we come to exact definitions of measures for success, let us provide some motivation for such a measure and link it to a debate which lasts now longer than two millennia, namely the epistemic debate about scepticism.

It is clear that nature counts as benchmark regarding success. However, regarding the aim of such a game between nature and individual i one might postulate different things: One aim might be to collectively score best: Since nature itself is the benchmark, it always has minimal loss: $\ell(y_t^s, y_t^s) = 0$ (due to axiom 2.1), and hence also maximal success: $s(y_t^s, y_t^s) = 1$. In order to increase the collective score, the best strategy to perform for nature would be to behave exactly like i predicts. To continue this metaphorical talk, given such an aim, nature has to behave nicely, has to perform a supportive and innocent strategy 🐦:

Definition 2.6 (Nature's Strategy: *Angel*)**.**

$$\text{🐦} : y_t^s \in \arg\max_{x\in[0,1]} s(f_{i,t}^s, x) \quad \forall s, t \in \mathbb{N}$$

$$E.g. : y_t^s = f_{i,t}^s$$

If nature plays such a strategy then $s(f_{i,t}^s, y_t^s) = 1$. It is clear that postulating such an aim for the epistemic realm is very optimistic, naïve, and innocent too. Concerning the interplay between perception and reality, e.g., *naïve realism* might be a paradigmatic case in point of a position which argues from a strategy like 🐦. However, this is no common position in epistemology. On the contrary, epistemologists are usually doubting such an assumption, especially for the epistemic realm. What is much more common is a sceptical perspective. The strongest version of such a scepticism is the assumption that nature plays against an individual, by a strategy 🐱 (*Daemon*) which tries to minimise the individual's success by maximising its loss. Most of the times such a sceptical perspective serves a methodological purpose: Starting with sceptical considerations, epistemologists often want to end up with reasons why such a sceptical position is untenable. In order to provide strong reasons based on a strong foundation, they assume a strong form of scepticism which tries to undermine any foundation whatsoever. Our application in part II of the book will be in a similar line. However, for now we want to have a short look on traditional sceptical positions in order to motivate our approach via meta-inductive prediction games.

If one considers, e.g., ancient scepticism, then one has to mention first and foremost Pyrrho, the founder of *Pyrrhonian scepticism*. However, since Pyrrhon himself seems to have intended to advocate sceptical lifestyle rather than to provide reasons for *suspension of judgement*, in the epistemic realm more relevant are his followers. Most prominent is perhaps Sextus Empiricus who distinguished three epistemic approaches: *dogmatism*, which claims "to have found the truth", *Academic scepticism*, which "asserts that it cannot be apprehended", and *general scepticism*, which is "still searching" the truth (see Sextus Empiricus 1999, book I, sect.1, p.45). Dogmatists like Aristotle might be interpreted as seeing things through rose-coloured glasses as discussed above: They seem to think that somehow they managed to bring about that nature plays strategy 🐦. General sceptics, like Sextus Empiricus, so it seems, are opposing dogmatists inasmuch as they think that assuming strategy 🐦 is not justified. However, Sextus Empiricus does, e.g., also not call into question the very existence of an external world. Rather, he suggests universal suspension of judgement. He claims, e.g.:

> "We do not reject the things that lead us involuntarily to assent in accord with a passively received *phantasia*, and these are appearances. [. . .] For example, the honey appears to us to be sweet. This we grant, for we sense the sweetness. But whether it is sweet we question insofar as this has to do with the [philosophical] theory, for that theory is not the appearance, but something said about the appearance. [. . . In general] nature's guidance is that by which we are naturally capable of sensation and thought; [. . .] hunger drives us to food and thirst makes us drink; [. . . Important is that] we say all these things without belief." (see Sextus Empiricus 1999, book I, sect.10, p.49)

So, what Sextus Empiricus seems to claim is that a prediction task as framed in terms of a game such as above is not a game he would join playing. Rather, he prefers to "say that as regards belief the Skeptic's goal is *ataraxia*" which "is an untroubled and tranquil condition of the soul" and the best means to achieve this goal is to perform *epochē* which "is a state of the intellect on account of which we neither deny nor affirm anything" (see Sextus Empiricus 1999, book I, sect.4 and 12, p.46 and p.49). Hence, he seems to suggest that any epistemic aim of a prediction game runs contra a practical aim, namely tranquillity, for which reason he thinks one should better abandon playing such games.

A sceptical position that fits more to our setting is the position of Arcesilaus who is said to be responsible for turning Plato's Academy to a specific branch of scepticism, the Academic scepticism. Although also Arcesilaus argued for suspension of judgement as a final mode of thought, contrary to Sextus Empiricus he thought it to be actually a very good means to join a game as described above. His idea was to engage into discussion in form of a debating contest (*logos* in the sense of discourse) in order to prove suspension of judgement resulting of debating. Even Socrates' slogan *to know that one does not know* turns in Arcesilaus' view to *not knowing* to know that one does not know (see Thorsrud 2009, pp.43f). And one might suspect that the iteration of *not knowing* goes on and on. However, methodologically he suggested to join discussion and by this, according to our framing, to join prediction games as indicated above. e.g. his critique of Stoic philosophy is summarised via reference to Marcus Tullius Cicero as follows:

> "For any sense-impression S [represented by our $f^{s}_{i,t}$], received by some observer [i], of some existing object O [represented by our y^{s}_{t} ...] we can imagine circumstances in which there is another sense-impression S' [our $f^{s'}_{i,t}$], which comes either (i) from something other than O [i.e. $y^{s'}_{t}$], or (ii) from something non-existent, and which is such that S' is indistinguishable from S to [i]. The first possibility (i) is illustrated by cases of indistinguishable twins, eggs, statues or imprints in wax made by the same ring. The second possibility (ii) is illustrated by the illusions of dreams and madness [that is strategy 🐱]." (see Thorsrud 2017, sect.2)

Anecdotal evidence via Diogenes Laërtius suggests that by help of such a debate Arcesilaus was able to trick a student of the founder of the Stoic school, Zeno of Citium, into thinking that wax pomegranates were real (see Thorsrud 2009, note 13 of chpt.3). Although this must have produced an awful taste experience for the poor student, whatever the practical consequences may have been, it is noteworthy that Arcesilaus was one of the first to prominently refer to illusions of dreams etc. in order to exploit the setting of a "prediction game" against any rationale of believing or disbelieving.

Such reference was prominently made especially in the modern era by methodological sceptics like Descartes or Hume. Descartes, e.g., suggests:

> "Accordingly, I will suppose not a supremely good God [i.e. our 🐦], the source of truth, but rather an evil genius [our 🐱], supremely powerful and clever, who has directed his entire effort at deceiving me. I will regard the heavens, the air, the earth, colors, shapes, sounds, and all external things as nothing but the bedeviling hoaxes of my dreams, with which he lays snares for my credulity.
>
> I will regard myself as not having hands, or eyes, or flesh, or blood, or any senses, but as nevertheless falsely believing that I possess all these things." (Descartes 1637/1998, par.22f)

And Hume formulates it this way:

> "All reasonings may be divided into two kinds, namely demonstrative reasoning, or that concerning relations of ideas, and moral reasoning, or that concerning matter of fact and existence. That there are no demonstrative arguments in the case, seems evident; since it implies no contradiction, that the course of nature may change, and that an object, seemingly like those which we have experienced, may be attended with different or contrary effects. May I not clearly and distinctly conceive, that a body, falling from the clouds, and which, in all other respects, resembles snow, has yet the taste of salt or feeling of fire? [our 🐱]" (Hume 1748/2007, sect.4, part ii, p.25)

We will discuss Hume's argument also in part II of the book in more detail. For now we want to end our excursus by mentioning the most prominent contemporary sceptical scenario debated in the literature, Putnam's thought experiment on a *brain in a vat*:

> "Imagine that a human being (you can imagine this to be yourself) has been subjected to an operation by an evil scientist [our 🐱]. The person's brain (your brain) has been removed from the body and placed in a vat of nutrients which keeps the brain alive. The nerve endings have been connected to a super-scientific computer which causes the person whose brain it is to have the illusion that everything is perfectly normal. There seem to be people, objects, the sky, etc; but really all the person (you) is experiencing is the result of electronic impulses travelling from the computer to the nerve endings." (Putnam 1981, pp.5f)

All these sceptical positions assume that the aim of playing a prediction game as outlined above consists not in nature playing *with* individual i (🐦), but in nature playing *against* i (🐱). At least for methodological reasons it is assumed that nature tries to minimise i's success. This is achieved by choosing y_t^s in such a way that it has most distance from $f_{i,t}^s$ within $[0,1]$ according to ℓ:

Definition 2.7 (Nature's Strategy: *Daemon*)**.**

$$\text{🐱} : y_t^s \in \arg\min_{x\in[0,1]} \mathscr{s}(f_{i,t}^s, x) \quad \forall s,t \in \mathbb{N}$$

Now, clearly 🐦 and 🐱 are not the only strategies possible. In a prediction game between a forecaster i with method F_i and nature, nature might play any strategy in between them, so also strategies according to which i's success lies on the interval:

$$\arg\min_{x\in[0,1]} \mathscr{s}(f_{i,t}^s, x) \ldots 0 \;\; \underset{\text{🐱}}{\vdash}\!\!\text{———————}\!\!\underset{\text{🐦}}{\dashv} \;\; 1 \ldots \arg\max_{x\in[0,1]} \mathscr{s}(f_{i,t}^s, x)$$

The upshot of methodological scepticism in our setting is that we have to allow nature to play any strategy within this interval, also the most sceptic one: 🐱. I.e., we are not allowed to make any assumptions (other than the boundary constraint on y_t^s of definition 2.4) about the event outcomes. We will see that there is a branch of machine learning which is concerned with prediction games as outlined above (definition 2.5) and which makes no assumptions about the event outcomes which have to be predicted, namely *online learning*. However, before we come to this branch, we give precise definitions of 'optimality' in the next section.

2.3 Success, Optimality, and Meta Predictors

When we introduced the notion of a prediction game and strategies 🐦 and 🐱 that marked a spectrum of strategies which nature can perform in such a game, we were also talking about a measure of loss ℓ and a simple measure for success $\mathscr{s}$. Now, these measures are about a single prediction. However, in online learning one considers not just one prediction, but a whole series of predictions (see figure 2.5) or a whole set of possible predictions. This broader perspective allows us also to define a broader set of measures for success.

If we do not want to explicitly refer to the true value series $\mathcal{Y}$, but rather leave it open to the context to either specify it further or allow for any such series, we will use an indexed notation for ℓ:

Definition 2.8 (Loss, Indexed)**.**

$$\ell_{i,t} = \ell(f_{i,t}, y_t)$$

In principle we have here a loss ℓ which is monotonically increasing with the distance between $f_{i,t}$ and y_t in mind. Also in most of the results below we assume a (in its first argument) convex loss function. Note that this definition and all definitions below are to be understood as definitions given some prediction game G with the series of the true values $\mathcal{Y}$ and a set of prediction methods $\mathcal{F}$. If we need to specify G further, we will mention it explicitly in the respective definition or theorem. If there is no need for a specification, we will often also refrain from mentioning G explicitly, as we did, e.g., already in definition 2.8.

What we called *simple success* in definition 2.3, is often called *score* in online learning. It is the value which a prediction method F_i earns for one prediction about an event's outcome at time or round t, namely the prediction $f_{i,t}$.

Definition 2.9 (Score)**.**

$$s_{i,t} = 1 - \ell_{i,t}$$

Note that the more fine-grained notion of the score defined on a specific type s of an event is $s^s_{i,t} = 1 - \ell(f^s_{i,t}, y^s_t)$. We will make use of this more fine-grained notion later on in the probabilistic setting.

Now, the score of a prediction method does not say much about its general success. What matters in evaluating a prediction method is how well it scores in general. For this reason we introduce a measure of absolute success which just states for each point in time or round t how a method scored in sum (see Schurz 2008b, p.279):

Definition 2.10 (Absolute Success)**.**

$$asucc_{i,t} = \sum_{u=1}^{t} s_{i,u}$$

More importantly for the optimality results later on is a measure of relative success which just states for each point in time or round t how a method scored on average (see the notion of the *success rate* in Schurz 2008b, p.279):

Definition 2.11 (Relative Success)**.**

$$succ_{i,t} = \frac{\sum\limits_{u=1}^{t} s_{i,u}}{t}$$

This kind of success is the most important one for the remainder of this book and when we speak of *success* without modification, then we have this notion in mind. The idea will be to optimise it in several epistemic applications. Also here the more fine-grained measure $succ^s_{i,t}$ can be defined in the same way, but on basis of $s^s_{i,t}$ instead of $s_{i,t}$.

For some more general cases we need also two further and somehow weaker notions of success. First, there is the notion of *average group success* which is just the average of the success rates of a group of prediction methods:

Definition 2.12 (Average Group Success)**.**

$$\overline{succ}_{\{i_1,\ldots,i_m\},t} = \frac{\sum\limits_{j=1}^{m} succ_{i_j,t}}{m}$$

The more fine-grained measure $\overline{succ}^s_{\{i_1,\ldots,i_m\},t}$ is, again, based on $succ^s_{i,t}$. Clearly, if the group consists of just one forecaster, then $succ = \overline{succ}$.

The second weaker notion of success is the notion of *expected success*. It covers the case where we do not have information about the exact forecast of a forecaster, but we know just the probabilities of her forecasting specific values. The common notion of an *expected value* is usually defined with respect to random variables: For a discrete random variable Z_1 with the value space $\mathcal{V} = \{v_1, \ldots, v_{\mathcal{k}}\}$ and the probability distribution Pr over $\mathcal{V}$, the expected value of Z_1, i.e. $\mathbb{E}[Z_1]$, is defined as:

Definition 2.13 (Expected Value)**.**

$$\mathbb{E}[Z_1] = \sum_{l=1}^{\mathcal{k}} Pr(Z_1 = v_l) \cdot v_l$$

If, e.g., $\mathcal{V} = \{0.0, 0.5, 1.0\}$, and $Pr(Z_1 = 0.0) = Pr(Z_1 = 0.5) = Pr(Z_1 = 1.0) = 1/3$, then $\mathbb{E}[Z_1] = 0.5$. In the same way we can define a measure for the expected success of a forecaster with method F_i by considering her probability of predicting the true value. By averaging along the rounds we get (see Shalev-Shwartz and Ben-David 2014, p.252):

Definition 2.14 (Expected Success).

$$\mathbb{E}[succ_{i,t}] = \frac{\sum_{u=1}^{t} Pr_i(f_{i,u} = y_u) \cdot s_{i,u}}{t}$$

where Pr_i is i's randomisation of her prediction
(for details see section 4.1)

The more fine-grained measure $\mathbb{E}[succ^s_{i,t}]$ is defined the same way, but on basis of $f^s_{i,t}$ and y^s_t.

These are the three main measures of success we will use for proving the general results employed in the book. For one or another application we will define further measures of success. However, they will be fitted very much to specific cases and they will be still based on these ones.

Now, the main aim of optimisation in a prediction game is to construct a prediction method F_m with predictions f_m whose success is optimal compared to all the other prediction methods of the game. As we have seen above, in the online learning paradigm with adversarial development the outcome can be always such that F_m's success is minimal. This is the case when nature plays strategy 🐱 against F_m. However, one has to be aware of the difference between *maximal*, *optimal*, *minimal*, and *suboptimal* success. Clearly, the maximum of $succ$ is 1. Its minimum is 0. From the sceptic's perspective it is important to note that nothing hinders nature to perform 🐱 and by this enforce minimal success for such a predictor F_m. However, things are different when we switch from *maximality* to *optimality*. Optimality and suboptimality are not absolute notions, but relative ones. In prediction games they are relative to the predictors with highest success. Nature might play a strategy which lets these predictors' success rate deviate from the maximal one. It might even play strategy 🐱 in such a way that the highest success of a player in a setting equals the minimum 0, so its strategy is super-adversarial. However, as we will see in chapter 3, there is an online learning algorithm or prediction method F_m such that nature's strategy can not be such that F_m is suboptimal, i.e. not optimal.

Optimality with respect to success consists in having the highest success compared to all other methods or agents in the setting. The degree of suboptimality consists in the deviation of being optimal, i.e., the deviation of the highest success of a prediction method in the game. Before we define a measure for optimality, we define a measure for the difference of the successes or losses of two players, regardless of whether one of them is optimal or not. Concerning the losses, this difference can be also interpreted as a degree of *regret* of one prediction method having not predicted the same way as the other: For this reason the difference of the accumulated loss of two forecasters is also called '*regret*' in the machine learning literature (see Cesa-Bianchi and Lugosi 2006, p.2). Since we assumed the loss measure to be within the unit interval (axiom 2.2) and we defined the simple success and score of a forecaster i as 1 minus its loss (definitions 2.3 and 2.9), we get (see Shalev-Shwartz and Ben-David 2014, p.251):

Definition 2.15 (Absolute Regret)**.**

$$aregret_{\langle i,j\rangle,t} = \sum_{u=1}^{t} \ell_{i,u} - \sum_{u=1}^{t} \ell_{j,u} = \sum_{u=1}^{t} (s_{j,u} - s_{i,u})$$

Again, the event variable-relative notion is $aregret^{s}_{\langle i,j\rangle,t}$ and based on $s^{s}_{i,t}$ and $s^{s}_{j,t}$. Note that $aregret$ up to time or round t is within the interval $[-t,t]$. If it is positive, then j's unaveraged net success is higher than that of i and so i regrets having performed a different strategy. If it is negative, then i's unaveraged net success is higher than that of j and so i does not regret having performed her strategy (while clearly j does in comparison to i). If it is zero, then they are on a par and there is no reason for regretting—for neither of them. Clearly $aregret_{\langle i,j\rangle,t} = -aregret_{\langle j,i\rangle,t}$.

For reasons of completeness, we also introduce the notion of relative regret which takes the average of the absolute regret up to round t:

Definition 2.16 (Relative Regret)**.**

$$regret_{\langle i,j\rangle,t} = \frac{\sum_{u=1}^{t} \ell_{i,u} - \sum_{u=1}^{t} \ell_{j,u}}{t} = \frac{\sum_{u=1}^{t} (s_{j,u} - s_{i,u})}{t}$$

Now, optimality seems to consist in having nothing to regret in the following sense: Those agents within a setting are optimal, who need not regret having performed a different strategy than the other ones, i.e. those is, whose regret with respect to all other agents j is not positive, i.e.: $aregret_{\langle i,j\rangle,t} \leq 0$. We could use this as a definition of 'optimality'. However, since we introduced different measures of success, we also want to define several notions of *optimality* based on the different success measures (the notion of *access optimality* was introduced by Schurz and Thorn 2016):

Definition 2.17 (Optimality)**.**
Forecaster i is access optimal in the long run in G iff for all $1 \leq j \leq n$:

$$\lim_{t\to\infty} (succ_{i,t} - succ_{j,t}) \geq 0$$

Forecaster **group** $\{i_1, \ldots, i_m\}$ is access optimal in the long run in G iff for all $1 \leq j \leq n$:

$$\lim_{t\to\infty} (\overline{succ}_{\{i_1,\ldots,i_m\},t} - succ_{j,t}) \geq 0$$

Forecaster i is **expected** to be access optimal in the long run in G iff for all $1 \leq j \leq n$:

$$\lim_{t\to\infty}(\mathbb{E}[succ_{i,t}] - succ_{j,t}) \geq 0$$

Note that the first notion is equivalent with demanding having no positive regret in the long run. We call these methods '*access* optimal', inasmuch as their being optimal presupposes that we have access to the method's of G success-rates, in particular access to their past predictions. And we call these methods 'optimal *in the long run*', inasmuch as we compare their successes in the limit.

One important property of optimality is that it is independent from any threshold of $succ$. So, if, e.g., the best prediction method fares worse than a method would fare that flips a fair coin ($\mathbb{E}[succ] = 0.5$ in the long run), this method would still turn out to be optimal. Even in case of $\max(succ_1, \ldots, succ_n) = 0$ in the long run, i.e. in a super-adversarial setting, all methods are optimal simply because *optimality* is defined in purely relative terms. A prediction method i is suboptimal, if it is not optimal. And i is *strictly* suboptimal, if it is outperformed by all other accessible prediction methods j in the sense that $\lim_{t\to\infty}(succ_{i,t} - succ_{j,t}) < 0$.

Up to now we have described measures for evaluating prediction methods. We also defined what it means for a prediction method to be access optimal in the long run. And we claimed that in online learning one finds algorithms which are guaranteed to be optimal. In what specific sense they are optimal will be characterised in chapter 3. But before we come to this, we need to introduce into our setting one more notion, namely the notion of a *meta predictor*.

All prediction methods we were talking up to now have one thing in common: They are methods that map information about past outcomes and some Input to their prediction:

$$F_i : \text{Input}(s,t) \longrightarrow f_i$$

We have indicated in section 2.2 that this Input might contain also, e.g., a priori principles which are used for hypothesis construction. Regarding the ingredients of a prediction game we distinguished prediction methods only according to the information they base their predictions on regarding $\mathcal{Y}$ (the outcomes): Clairvoyants have access to future outcomes, oracles to outcomes of strangely related types of events, abductive forecasters have access to more systematically related types of events etc. However, the setting of a prediction game allows also to employ some further information. Besides $\mathcal{Y}$, there is also $\mathcal{F}$ (the predictions) which might be used for constructing a prediction. So, in principle it is possible to characterise *meta methods* F_m that are defined not only on $\mathcal{Y}$ and some Input, but also on the *non*-meta methods'—sometimes also called 'candidate methods' (see Schurz and Thorn 2016)—predictions: $\mathcal{F}$:

$$F_m : \mathcal{F}, \text{Input}(s,t) \longrightarrow f_i$$

Such meta methods can take on different forms: There is, e.g., the possibility to create hypotheses about the data based on the data and modify the hypotheses according to the other predictions. Such a *mixed strategy* could be, e.g., object-induction plus normalising the prediction within the interval of the other predictions. However, for our purpose most important is the so-called *meta-inductive strategy* which performs induction not on the level of the data (like object-induction), but on the level of the predictions by performing induction on the success rates. We will see in the next chapter that this allows for defining an algorithm that is guaranteed to be access optimal in the long run. Hence its importance. Furthermore, there is the possibility to construct predictions not only out of data and candidate methods, but also out of meta methods. Such *meta-meta methods*

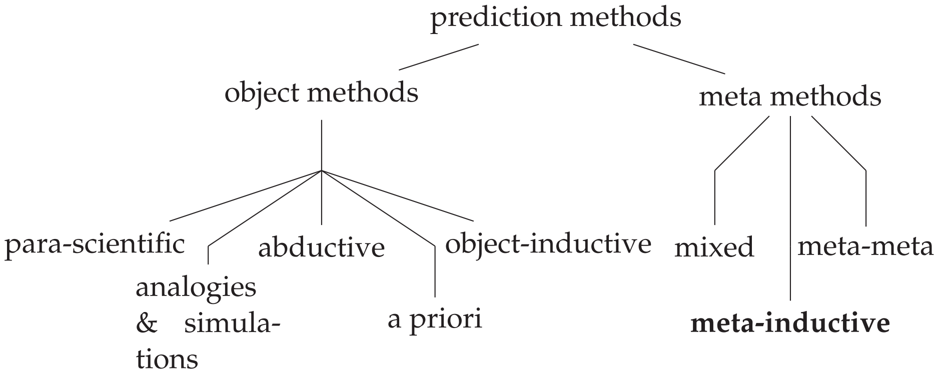

Figure 2.4: A taxonomy of prediction methods: Prediction methods can be divided into object methods and meta methods. Object methods base their predictions at most on information about events, but not information about other predictions. If their information basis is solely about past events of the same type, they are object-inductive. If it is also about past events of other types, they might be abductive or methods employing analogies and simulations. If they (claim) to have as information basis outcomes of events (of the same or different type) of the future, they are para-scientific. If they base their prediction neither on information about other predictions, nor on information about the events, they are a priori. Meta methods base their predictions at least on some information about predictions of other methods. If they make their predictions on the basis of such information and information about the events in question, they are mixed. If their predictions are purely on the basis of information about object methods' predictions, then they are meta-inductive. If they base their information solely on predictions of meta-inductive methods, then they are meta-meta methods (and so forth).

will turn out to be relevant regarding optimality results in a discrete setting, i.e. prediction games where the predictions are not within $[0, 1]$, but within a discrete set of values. Figure 2.4 provides an overview of the most important methods relevant for our endeavour.

Given a precise characterisation of ‘optimality’, we now want to search for prediction methods that allow us to achieve optimality. As we have mentioned above, there is a branch of machine learning which studies exactly such methods. In the next section we are going to indicate which branch is most relevant for our aim, namely the theory of meta-induction or online learning.

2.4 Machine Learning and Methodological Scepticism

A great deal of the machine learning literature focuses on the task of designing and studying algorithms for making predictions in a prediction game as characterised in the preceding section. Learning consists of finding a hypothesis “in” some data, or better: constructing a prediction method F_i based on the data $\mathcal{Y}$. It is machine learning, inasmuch as such constructions should be performable also by machines.

One can differentiate several types of *learning*, depending on the following four parameters (see Shalev-Shwartz and Ben-David 2014, sect.1.3):

1. *supervised vs. unsupervised*: Learning is supervised, if the data provided for learning ($\mathcal{Y}$) contains information about whether the property, which should be learned, applies to the data or not. So, e.g., if one is supposed to predict whether it will rain or not and if the data provided for learning allows the learner to distinguish rainy days from non-rainy ones and identify the rainy ones, then the learning process is supervised. On the other hand, if the learner gains only information about differences in the data without her being able to identify rainy days, then learning happens unsupervised.

2. *active vs. passive*: If the learner can intervene on the data set presented to her, then learning is active. If not, then it is passive. In the first case the learner can (systematically) decide, e.g., for which parameters data should be generated. In the latter case the learner has to take what she gets as input.

3. *non-adversarial vs. adversarial*: If the data is selected such that it contains *prototypical* cases, then it is non-adversarial. If it is generated in a *random* way, e.g., then it is indifferent (typically *big* random data is also *valuable* and by this non-adversarial). If it is generated in a

way that is even negatively correlated to finding the right hypothesis/making the right predictions by the learner, i.e. if the environment is *adversarial* to some degree, then it is not valuable.

4. *sample based vs. online*: If the learner gets a data set (learning phase) before she has to make a prediction (prediction phase), then she has a so-called *batch learning protocol*. If she has to respond *online*, i.e. throughout the learning process (prediction phase = learning phase), then she has a so-called *online learning protocol*. In the online case the learner may become an expert over time, whereas in the sample-based case the learner might become an expert already before she makes any predictions.

We call the first parameter of each pair a *positive parameter*, for if it is assumed to hold, then the learner has some further possibilities to act or grasp information.

For illustrative purposes, the difference between online and sample-based learning is schematically represented in figure 2.5: In online learning one has to make a prediction about an event in question, receives afterwards the outcome, and can use this to learn a hypothesis for a prediction about the next event. In contrast to this, sample-based learning usually consists of a big data set which is exhausted in a learning phase. Then, after constructing a hypothesis about the data, one starts with the prediction phase. Note that results about sample-based learning very often require an assumption about a distribution of random variables representing the events: If Z_1 and Z_2 are random variables representing two specific, but arbitrarily chosen, events of an infinite event series, then they must be independent and identically distributed (*i.i.d.*):

Definition 2.18 (I.I.D.)**.** An event series represented by random variables is an *i.i.d.* series iff for any two events of the series represented by Z_1 and Z_2 it holds:

- *Identical distribution*: $\forall x \in [0,1] : \; Pr(x \geq Z_1) = Pr(x \geq Z_2)$
- *Independent distr.*: $\forall x \in [0,1] : \; Pr(x \geq Z_1 | x \geq Z_2) = Pr(x \geq Z_1)$

Intuitively, this condition states that the data set used in sample-based learning "is a window through which the learner gets partial information about the distribution [. . .]. The larger the sample gets, the more likely it is to reflect more accurately the distribution and labelling used to generate it" (see Shalev-Shwartz and Ben-David 2014, p.18). For sample-based learning this is a crucial assumption: It is necessary for proving optimal learning performance. We will see in chapter 3 that in online learning, at least in the continuous case, no such assumption is needed. Note that our setting above (definitions 2.4 to 2.5) is, up to now, much simpler and considers only one value of an event. So, random variables are not used by us right now. However, later on when we introduce probabilistic prediction games we will show a way of implementing random variables and transforming the optimality results to this kind of games.

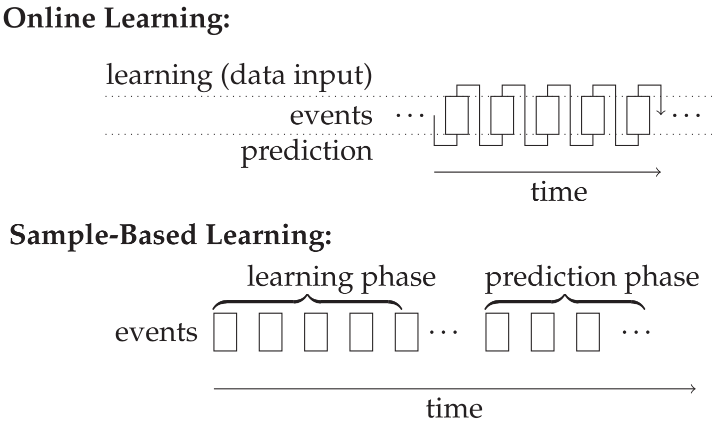

Figure 2.5: Difference between *online learning* and *sample-based learning*: In online learning the predictor (i) provides a prediction of an event, (ii) receives information about the outcome which she can use as learning basis for (iii) providing her prediction of the next event etc. In sample-based learning the predictor (i) receives a set of data (information about the outcome of a series of events) which she can use for (ii) providing her predictions on another series of events.

Before we provide some motivation for the *learning paradigm* which we are going to employ in this book, let us give some examples: *Unsupervised* learning is learning where the items of the data set whose pattern should be learned are not labelled. Usually the task of such learning is to identify items which do not conform to an expected pattern or other items in the data set. Such anomalous items (also called *outliers*, *noise*, *deviations*, and *exceptions*) are interpreted then as an error, a problem or a structural defect. Applications of this kind of learning are manyfold: It is used in image analysis, pharmaceutical research (e.g. for finding novel molecular structures), for detecting mislabelled data in a training set, etc. (see Hodge and Austin 2004).

As a prototypical example of *supervised* learning one might consider the task of a (natural) scientist which is very often to generate a theory out of a set of data with clearly distinguished phenomena and performing experiments for validating the theory. According to the parameters above, this task falls under the following learning paradigm: *Supervised active sample-based learning based on valuable/indifferent/not valuable data*. It is *supervised* inasmuch the data set scientists work with is usually operationally accessible in the sense that they are able to identify the relevant phenomena therein. A psychologist, e.g., who theorises about a correlation between frustration, aggression, and depression usually lists a set of operational properties for each of these phenomena that allows her to clearly distinguish them. Such learning is *active* inasmuch as scientists ideally, whenever possible, try to perform experiments in a controlled setting which allows them to state clear conditions for intervention. By this they do not passively grasp data, but generate it. Furthermore, in most of the cases it is *learning by help of a sample* in the sense that in most of the cases one starts already with a given data set or one generates data on which one bases her working hypothesis. Whether a hypothesis generated out of the data is true and whether by this the data was *prototypical* (also: *externally valid*) or *random* regarding the scientist's task of finding a true hypothesis, or whether it was *adversarial* with respect to her task, is left open here and discussed in part II since this question concerns the fundamental epistemic problem of induction.

A more determinate example within the same paradigm is learning by a student in a lab, where usually the data provided by the instructor is *prototypical*. So this kind of learning is within the paradigm of *supervised active non-adversarial sample-based learning*.

An example for the paradigm of *learning according to an online protocol* is the case of a stockbroker who has to make every day a decision or prediction which is based on her experience gathered so far. Different from *learning by help of a sample* the learner has to present her predictions "on-line". In general, her way of learning is *supervised* inasmuch as she gets feedback about the true outcome of a stock value. It is also passive inasmuch as there is usually little space for interventions since the setting is not a controllable one (foreseeing of interventions of some stockbrokers who are so influential that in many cases their hypothesis becomes a *self-fulfilling prophecy*). Although there are lots of mistakes and errors in predictions, one usually assumes that the data used is not *adversarial* with respect to the predictions.

Finally, an example of the paradigm of *supervised passive adversarial online learning* is learning of a spam filter: Here usually the filter learns in a *supervised* way since messages are labelled as spam or not-spam. It is typically *passive*, since the filter has to wait until a user tags a message. It is *online* since a message is sent to the filter, the filter has to predict whether it is spam or not, in case it is not spam it is directly forwarded to the user who, ideally, provides information about whether it was spam or not. Furthermore, it is usually *adversarial* since senders of a message that is spam aim at a negative correlation between the filter's prediction that the message is not spam while it *de facto* is spam (see Shalev-Shwartz and Ben-David 2014, p.5).

We now come to a motivation for selecting the *learning paradigm* which we employ in this book for epistemic purposes, namely the *adversarial su-*

pervised passive online learning paradigm. In table 2.1 the different learning paradigms with examples are listed. As we have described above, a successful scientist's way of learning is most demanding regarding the parameters: She needs time for a training phase (*sample-based*), the possibility to experiment (*active*), all the data to be labelled (*supervised*), and luck to be not in a daemonic setting (*non-adversarial*). Assuming that we *de facto* learn this way is, epistemically speaking, naïve in the sense that one assumes nature to play a supportive strategy 🐦. The extreme on the other side is *unlucky guessing* where one has to make a prediction on-line, i.e. has little data available, cannot intervene on the data, cannot identify the relevant data for learning since it is unlabelled and is even in a daemonic setting. Due to the lack of any "positive information" we think it is problematic to even call this paradigm a *learning paradigm*. We suggest to speak of a *learning* paradigm only if at least one piece of "positive information" is available, i.e. if one gets to know the true results or labels of the dataset (*supervised*), one can perform experiments and intervene on the data (*active*), or one has at least a big enough dataset in order to learn some structure (*sample-based*). By this *unlucky* as well as *lucky guessing* drop out from the list. Furthermore, intervention with an online learning protocol seems to

non-adversarial	supervised	active	sample-based	example
0	0	0	0	Unlucky guessing
0	0	0	1	Unsuccessful ordinary anomaly detection
0	0	1	0	
0	0	1	1	Unsuccessful interactive anomaly detection
0	1	0	0	Spam detection
0	1	0	1	Ordinary data mining
0	1	1	0	
0	1	1	1	Anti-realistic ordinary science or student learning "by help" of a hostile instructor in a lab
1	0	0	0	Lucky guessing
1	0	0	1	Successful ordinary anomaly detection
1	0	1	0	
1	0	1	1	Successful interactive anomaly detection
1	1	0	0	Stockbroker
1	1	0	1	Ordinary data mining
1	1	1	0	
1	1	1	1	Realistic ordinary science or student learning by help of an instructor in a lab

Table 2.1: Examples for different (learning) paradigms. The dark grey coloured rows represent possibilities of parametrisation which we consider to be no *learning* paradigms. The white rows are those learning paradigms that are applicable in the two most sceptic scenarios. They are the *anomaly detection* and the *spam detection* paradigm.

provide a basis too small in order to allow for learning at all. If time or the number of rounds compensates for this, then online learning can be reformulated as some kind of sample-based learning in the sense that a learning algorithm based on an online learning protocol can be transformed to one based on a batch protocol (see Shalev-Shwartz and Ben-David 2014, exercise 21.5). So, we can foresee from this paradigm too. Hence, as table 2.1 shows, the two most sceptic (i.e. *adversarial*) settings where only one "positive" learning parameter is satisfied are the ones with the following prototypical applications: *ordinary anomaly detection* and *spam detection*.

Anomaly detection falls within the *sample-based learning paradigm*. As we have mentioned above, anomaly detection seeks, in a first phase, for anomalous items which are, in a second phase, interpreted as some kind of error or defect. Whereas learning (i.e. the first phase) in this paradigm is *unsupervised*, utilising the learned results (i.e. the second phase) needs some interpretation of the data which corresponds to labelling. So, although the learning process itself is *unsupervised*, the evaluation of what is learned falls within a *supervised* paradigm. Also that most of the optimality results on sample-based learning depend on an assumption about the structure of the events (i.i.d. as described in definition 2.18) makes this kind of learning richer in presuppositions (if it were really completely unsupervised, we would hesitate to call it a form of *learning*).

Spam detection, on the other side, falls within the *online learning paradigm*. Clearly, also here one needs *supervision* for evaluating what was learned. However, supervision is the only "positive" parameter assumed in this paradigm. So, at least it seems so, from an epistemic perspective the online learning paradigm of spam detection is *that* paradigm which can be applied in one of the most sceptic scenarios: If one has a basis for learning at all, i.e. if at least one positive parameter is assumed, then the sceptic's hardest challenge is to justify learning in an *adversarial supervised passive online paradigm*.

So, our motivation for seeking a solution to the sceptic's challenge in results of *online learning* is based on the fact that these results seem to be very parsimonious regarding presuppositions about the parameters of the framework. By this they appear to be promising for addressing a strong form of scepticism. As we will see in part II, we can employ online learning results even against a super-adversarial nature performing strategy 🐱.

Here also a terminological note is in order. In our investigation we refer to the theory for the outlined learning paradigm by the term *meta-induction* as well as *online learning theory*. Although both theories were developed in different disciplines—the former in philosophy and the latter in computer science—they both aim at designing prediction methods for the online learning paradigm for which reason reason we use the terms interchangeably.

To put the result of this section in a nutshell: We found a branch of *machine learning* which is concerned with an adversarial environment that has an analogue in traditional epistemology: *online* machine learning. Descartes' *daemon* found its way into the digital era: *Spam*.

Since we have all the ingredients needed, this is also the point from which we want to start with our survey of *epistemic optimality*: Online prediction games with meta-inductive forecasters. In the next chapter we are introducing the main result regarding epistemic optimality which underlies almost all applications of this book.

Chapter 3

The Logic of Deceivability

In this chapter further relevant distinctions of the theory of meta-induction and online learning regarding different learning tasks are explained: online classification and online regression. A general characterisation of the notion of online learnability *is provided. Afterwards, the logic of deceivability is introduced and a simple example of an online learning algorithm which is optimal, given there is a best expert in the setting, is provided. It is shown that by relaxing this assumption, the classification task can be solved only suboptimally. The regression task, on the other hand, can cope with this relaxation and proves to be optimal. This is the main optimality result exploit in the remainder of this book.*

Let us start with a coarse characterisation of the general notion of *learning*. As Shalev-Shwartz and Ben-David (2014, p.1) put it:

> "Roughly speaking, learning is the process of converting experience into expertise or knowledge. The input to a [learner] is training data, representing experience, and the output is some expertise, which usually takes the form of [some disposition to] perform some task."

This characterisation of learning is dispositional and in this sense *behaviouristic*—we do not talk about increased *understanding* etc., rather we simply speak of better performance. Think, e.g., on the task of learning a regularity of the form $\forall x(Px \rightarrow Qx)$ (stating, e.g., that all ravens are black). Initially, one might react to presented P-states in 50% with Q-answers and in 50% with $\neg Q$-answers. However, as soon as one performs better in the sense that one answers, e.g., in 70% of presented P-states with Q-answers, we would conclude that one has also learned the regularity a little bit better. We would conclude that one has fully learned the regularity, once one provides in 100% of presented P-states Q-answers. This dispositional notion of learning covers a wide range of phenomena: plants "learn" in the sense that they, e.g., react to stress factors in a way which increases their chance of "life prolongation". Animals as, e.g., rats learn in the sense that they avoid poisonous baits—when they encounter food with novel look or smell, they first try very small portions of it and if it causes illness, the look and smell is associated with the illness and rats will no longer take in such food. Clearly, also humans learn, e.g., when students are able to solve more exercises after than before attending a course. And the dispositional notion of learning is even so wide that it covers also learning of all other kinds of objects. Whether it makes sense to attribute learning to objects depends on our purposes and the features we want to describe. In the case of machines, as we have outlined above and we will see below, it definitely makes sense.

In section 2.4 we already distinguished sample-based learning from online learning. Given the dispositional learning paradigm we presented here, the difference between both can be described also in terms of *experience to expertise conversion*: In the sample-based learning case, a machine gains a huge data sample first, i.e. has lots of experience, and then starts to systematise the experienced facts in such a way that it ends up with an expertise algorithm that performs well. Whereas in the online case, a machine gains a little bit experience, tries to systematise in form of producing a little bit better expertise algorithm, gains further experience, and tries to end up with an even better expertise algorithm, and so on. We argued that online learning covers much more adversarial cases than sample-based learning does, since, first of all, an online learning algorithm receives much less input than a sample-based learning algorithm does, and secondly, sample-based algorithms are usually designed for cases where event outcomes of the sample are independently and identically distributed (*i.i.d.* see definition 2.18) with respect to the whole domain, whereas online learning algorithms are typically designed without any such assumption. For our purpose of application to epistemology, online learning covers better the sceptical case, since "in the online learning model we make no statistical assumptions regarding the origin of the sequence of examples. The sequence is allowed to be deterministic, stochastic, or even adversarially adaptive to the learner's own behavior (as in the case of spam e-mail filtering [or an *evil daemon*])" (see Shalev-Shwartz and Ben-David 2014, p.246). Also the learning target differs: In sample-based learning the goal is to learn something with a small expected loss or generalisation error, so *true* predictions are still the main aim (and it is achieved only by making assumptions about predictability properties of the sample). In contrast to this, the objective in online learning is to provide *optimal* predictions which is to minimise the regret (i.e. the difference between the cumulative loss of the algorithm and that of the experts in hindsight—see definition 2.15).

Up to now we have indicated only what it means that someone learns: By help of experience one gains expertise with respect to a specific learning target (e.g. with respect to true or optimal predictions). This was a learner-oriented perspective. However, one might also take in a problem-oriented perspective and wonder what characteristics a learning target or problem must have, in order to be learnable in general. This concerns the conditions for the learnability of a learning target or problem. As we will show in chapter 5, the traditional learning target put forward for epistemic justification (in particular induction), does not allow for learnability. So, given a traditional learning target for solving the problem of justifying induction, we cannot justify it. However, as we will also see there, putting forward as learning target optimality constraints of epistemic engineering as outlined in section 1.4, one can provide a justification of induction.

In this chapter we provide the most relevant results for this endeavour. Thereby we follow Shalev-Shwartz and Ben-David (2014), particularly chpt.21. We give a formal characterisation of the notion of *learnability* and distinguish two sub-tasks of online learning which have different learnability properties, namely online classification and online regression (section 3.1). Afterwards, we provide an introductory solution to the learnability problem of providing optimal predictions in case there is a best competitor accessible (section 3.2). We then show that there is no such solution for an online classification task, if one relaxes this condition. In order to do so, we provide a meta-inductive description of the *logic of deceivability* (section 3.3). Finally, we prove the main optimality result by showing that there is such a solution for an online regression task (section 3.4).

3.1 Online Learning and Learnability: Classification vs. Regression

In machine learning several learning paradigms are investigated. In section 2.4 we provided a fine-grained categorisation, which was formally correct in the sense of being complete and mutually disjunct regarding the binary parameters *non-adversarial/adversarial*, *supervised/unsupervised*, *active/passive*, and *sample-based/online* (for an overview see table 2.1 in this section). We have also argued that for the aim of epistemic engineering one best focuses on the *adversarial supervised passive online learning paradigm*, i.e. the learning paradigm according to which the environment provides values that are intended to minimise the learner's success (adversarial), the learner receives information about what were the true outcomes (supervised), the learner cannot perform experiments in the sense that she cannot ask for specific data (passive), and the learner receives the data only piecewise and has to make her predictions always before receiving data (online). This paradigm can be relevantly differentiated even further by help of a fifth parameter, namely the so-called *label-type* which concerns the values one is allowed to use in one's predictions (see Shalev-Shwartz and Ben-David 2014, pp.25f):

- *Classification*: In the case of classification, the data or events have to be labelled in a discrete way in the sense that they are assigned to a class, for example spam/non-spam. The task of the learner is to assign such discrete labels to new unlabelled data or events. For our setting this means that the values of the predictions or forecasts are in $\{n_1 \in [0,1], \ldots, n_m \in [0,1]\}$ (for some $n_1, \ldots, n_m$).
- *Regression*: In the case of regression, the data can be labelled in a continuous way: The learner has to provide for each event a label or value of $[0,1]$.

Note that in principle the *supervisor* could provide labels of a different type—but relevant for the categorisation in classification and regression is which labels the *learner* is allowed to use. Note further that in the machine learning literature often next to classification and regression, also *clustering* is mentioned. In the clustering case data is not labelled at all, but can be divided into groups based on similarity and other measures of natural structure in the data. An example would be the task of organising photos by faces without names, where one has to assign names to classes of photos. However, as we have argued in section 2.4, this would amount to unsupervised learning. In the sample-based case this concerns anomaly detection that is still in need of any labelling in the background, since otherwise one would not be able to speak of *success*. And in the online case this would amount to unlucky guessing, a situation which can be hardly understood as a case of *learning* (since there is no positive parameter present at all).

The main difference between classification and regression is that classification consists in discrete predictions, and regression consists in continuous (non-discrete) predictions. We can also make this distinction on the basis of our definition of prediction games (definition 2.5):

Definition 3.1 (Classification Game)**.** G is a classification game iff G is a prediction game and the predicted values in G are discrete, i.e.: There are $n_1 \in [0,1], \ldots, n_m \in [0,1]$ such that all $f^s_{i,t}$ of G are in $\{n_1, \ldots, n_m\}$.

Definition 3.2 (Regression Game)**.** G is a regression game iff G is a prediction game, but not a classification game.

As we will see soon, both kinds of games differ in important respects as they allow for different learnability properties.

Now, what exactly is the learnability property we are talking about? In order to motivate this notion, let us re-interpret the prediction setting we have introduced in chapter 2 for a moment (we concentrate on prediction games with a single event-type now): Such a prediction game consists of a sequence $\mathcal{Y}$ of the true event outcomes $y_1, y_2, \ldots$ and a set $\mathcal{F}$ of n sequences of predictions or forecasts $f_1 : f_{1,1}, f_{1,2}, \ldots$ and $f_2 : f_{2,1}, f_{2,2}, \ldots$

and ... and $f_n : f_{n,1}, f_{n,2}, \ldots$. We were interpreting the fs as the predictions of competing prediction methods, for which reason we were speaking of *prediction games*. Now, one could also interpret the fs not as predictions of competing prediction methods, but as *hypotheses* about the truth $\mathcal{Y}$. In this sense $f_1, f_2, \ldots, f_n$ are alternative hypotheses about the truth $\mathcal{Y}$ and $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ is the set of alternative hypotheses about the truth $\mathcal{Y}$. This interpretation corresponds better to the learning case. Now, to (absolutely) learn the truth on the basis of the hypotheses set $\mathcal{F}$ means that one becomes more and more an expert and in the end learns the truth. Strictly speaking, this would mean that one generates a hypothesis f_l (the learner's hypothesis) which predicts, on average, better and better, i.e. whose absolute success grows *strictly superadditive* with the number of prediction rounds t (formally: $succ_{l,t_1+t_2} \cdot (t_1 + t_2) > succ_{l,t_1} \cdot t_1 + succ_{l,t_2} \cdot t_2$). However, learning typically allows also for making errors, for which reason the process of learning might have also some "valleys", i.e. absolute success might grow not strictly additive with t (this feature of absolute success is a sufficient, but not a necessary condition for absolute learnability). Rather, a problem, the truth $\mathcal{Y}$, is online learnable on the basis of a set of hypotheses $\mathcal{F}$ in this absolute sense if there is an algorithm f_l which is guaranteed to reach $\mathcal{Y}$, at least in the long run, i.e. in the limit.

Now, clearly we cannot hope to find such an algorithm f_l which fits all prediction games. In an adversarial case, the truth $\mathcal{Y}$ of G is not learnable via any hypothesis set $\mathcal{F}$ in the absolute sense, since for any learning algorithm, i.e. prediction method or hypothesis, f_l which is non-parascientific, adversarial $\mathcal{Y}$ is simply defined as $y_t = 1 - f_{l,t}$. Hence, f_l receives at each round suboptimal score $s_{l,t} \leq 0.5$, and hence $succ_{l,t} \leq \frac{t \cdot 0.5}{t} = 0.5$. So $\lim\limits_{t \to \infty} succ_{l,t} \leq 0.5$ which implies that $\mathcal{Y}$ of G is not online learnable (in an absolute sense). Hence, in order to make sense of the notion of *absolute online learnability*, one needs to restrict the set of prediction games under consideration. The restriction which is relevant for online learnability (in an absolute sense) is that the truth $\mathcal{Y}$ is in the set of prediction methods or hypotheses: $\mathcal{Y} \in \mathcal{F}$ (in our later applications we assume furthermore that the "truth" is observable or accessible). In the machine learning literature this case is called *realisable*, since one can *realise*, achieve the absolute learning target (see Mohri, Rostamizadeh, and Talwalkar 2012, p.150; and Shalev-Shwartz and Ben-David 2014, sect.21.1). We define:

Definition 3.3 (Realisable Game). A prediction game G with $\mathcal{Y}$ and $\mathcal{F}$ is weakly realisable iff there is an $f_i \in \mathcal{F}$, and a $u \in \mathbb{N}$ such that for all $t > u$: $f_{i,t} = y_t$.
We say that G is *strictly realisable* or simply *realisable*, if $u = 0$ and there is *exactly one* such $f_i \in \mathcal{F}$. For short we will also write for this case: '$\mathcal{Y} \in \mathcal{F}$'.

So, in principle, for learning the truth it suffices to include into $\mathcal{F}$ any hypothesis which deviates from the truth $\mathcal{Y}$ only in finitely many instances. However, the common notion of *(strict) realisability* as defined above allows for simpler theorems about the short run performance of algorithms and regarding the long run properties there is no difference. So, if not stated otherwise, we use the *strict* notion of *realisability*.

Although, as we will see later on, *realisability* is too strong an assumption in order to be cashed out for approaching the problem of induction, we want to highlight that the framework of realisable prediction games might be relevant for studying *a logic of deceivability* with respect to externalist accounts of epistemic justification (see our discussion of externalism in section 1.1 as well as Grundmann 2009, 2017). The idea in a nutshell: Externalism allows for an evaluation of epistemic attitudes of an epistemic agent by help of means which are not accessible to the (or any other) epistemic agent. So, e.g., in *Gettier cases* or *fake barn cases* the agent lacks knowledge because her method of forming a belief was not reliable, although there might be no way for her to distinguish between a reliable and an unreliable method for forming such beliefs. The notion of *justification* and *reliability* is external and, although there are no clear (internal) rules for applying them, there (external) rules: Given the external description of such cases we can distinguish between reliable/unreliable methods, externally justified/unjustified beliefs. Now, similarly, an epistemic agent might not be aware of whether she is part of a realisable prediction game or not (i.e. this is not internal to her). But from an external perspective, knowing that she is, one might wonder whether her learning algorithm allows also for learning the truth. In this sense realisable prediction games might provide an interesting framework for studying an externalist notion of *reliability* (I would like to thank Gerhard Schurz for pointing this out to me).

With this idea of realisable games at the back of our mind, we define the notion of *absolute online learnability* as follows (this is a modified version of “online learnability” as characterised by Shalev-Shwartz and Ben-David 2014, p.246, dfn.21.1):

Definition 3.4 (Absolute Learnability)**.** Given some condition C, a hypothesis set $\mathcal{F}$ allows for online learnability (in an absolute sense) iff there is an $f_l \notin \mathcal{F}$ such that for all prediction games G with $\mathcal{F} \cup \{f_l\}$ and $\mathcal{Y}$ that satisfy C it holds: f_l is not para-scientific/no clairvoyant (i.e. based on $\mathcal{Y}$), and:

$$\lim_{t \to \infty} succ_{l,t} = 1$$

If f_l is an algorithm in the sense of above, we will also say that $\mathcal{F}$ is online learnable in the absolute sense by help of f_l or that f_l allows for absolute online learnability of $\mathcal{F}$.

We will also simply say that $\mathcal{F}$ *is absolutely online learnable*, although what is meant is that $\mathcal{F}$ allows for learning the truth $\mathcal{Y}$. The condition that

f_l is not based on $\mathcal{Y}$ is needed for ruling out trivial cases of learnability, e.g., expansions of the method or hypothesis set $\mathcal{F}$ by $\mathcal{Y}$ itself. Again, details are a little bit more sophisticated and what is meant, strictly speaking, is that the definition of $f_{l,t}$ is not based on $y_{\geq t} \in \mathcal{Y}$—i.e., that f_l is no para-scientific method according to our characterisation in section 2.3 (see also figure 2.4). Clearly, if there is an algorithm f_l which allows for absolute online learnability, this means that f_l is guaranteed to find the truth $\mathcal{Y}$ in the following sense: $\mathit{succ}_l = 1$ in the long run only, if f_l scores maximally in the long run, i.e. if f_l's score $\mathfrak{s}_l = 1$ in the long run (via definition 2.11). But this means that f_l's loss vanishes in the long run: $\ell(f_l, y) = 0$ (via definition 2.9). Trivially, also the truth's loss "vanishes": $\ell(y, y) = 0$ (via axiom 2.2). So, if we were to partition the set of hypotheses and the truth according to their long run losses, we could not differentiate the learner and the truth $f_l =_\ell y$. So, a prediction game with such an f_l contains the truth and hence is realisable. In the following sense is *realisability* a necessary condition for *absolute learnability*:

Corollary 3.5 (Absolute Learnability ⇒ Realisability)**.** *If $\mathcal{F}$ of a prediction game with truth $\mathcal{Y}$ is online learnable in an absolute sense by help of f_l, then a prediction game G' with $\mathcal{F} \cup \{f_l\}$ and $\mathcal{Y}$ is realisable.*

Note, however, that realisability of G itself is not necessary for absolute learnability: Assume, e.g., a prediction game G with $\mathcal{F} = \{f_1, f_2\}$ such that $f_{1,t} = f_{2,t} = 1 = 1 - y_t$ for all $t \in \mathbb{N}$. So, G is not realisable. Now, an algorithm f_l defined on $\mathcal{F}$ as $f_{l,t} = min(0, f_{1,t} - f_{2,t})$ is not para-scientific (its definition does not include y at all), and it online learns in G via $\mathcal{F}$ absolutely the truth, since for all t: $\mathit{succ}_{l,t} = 1$. So, success is not excluded for such an f_l in the not realisable case. However, a deceiver can always devise a game G'' according to which f_l is maximally unsuccessful, simply by defining $y_t = 1$. Then for all t: $\mathit{succ}_{l,t} = 0$ and hence f_l never learns the truth via $\mathcal{F}$. So, there is no guarantee for the algorithm to reach the truth, if the game is not realisable.

Now, as we will see soon, there are learning algorithms f_ls which allow for absolute learnability under the realisability and a further characteristic condition of the hypothesis set. So, if we specify C to this further characteristic condition, then *realisability* is also sufficient for absolute online learnability. As we will show in part II (chapter 5) in detail, the classical conditions for solving the problem of epistemic justification rule out the realisability of such games, hence, according to the classical constraints truth cannot be learned in absolute terms. This impossibility result asks for another approach to the problem of epistemic justification, namely that of relative learnability, or optimality. To (relatively) learn means not that one becomes more and more an expert (in the absolute sense) and in the end learns the truth. Rather, it means that one becomes more and more and expert in the relative sense that one is less and less outperformed by the other prediction methods or hypotheses and becomes the best predictor or hypothesis in the setting. In section 2.3 we defined notions which capture this quite well, namely the notions of *optimality* and *regret*. Recall, the latter consists in the difference between one's own cumulative loss and that of a competitor method or hypothesis (see definition 2.15). Positive regret with respect to a prediction method or hypothesis means that the predictive success of that method or hypothesis is greater than one's own. Negative regret means that one's own predictive success is greater than that of the other prediction method or hypothesis. That one is less and less outperformed by one's competitor methods means that the regrets with respect to the competitor methods or hypotheses grow only *strictly subadditively* with the number of rounds t (formally: $aregret_{\langle l,i\rangle,t_1+t_2} < aregret_{\langle l,i\rangle,t_1} + aregret_{\langle l,i\rangle,t_2}$ for all $1 \leq i \leq n$). Again, also relative learning might have "valleys"; hence, strict subadditive growth of the regret with t is a sufficient condition for relative learnability, but not a necessary one. Rather, a problem, the hypothesis set, $\mathcal{F}$ is online learnable in this relative sense if there is an algorithm f_l which is guaranteed to have no (positive) regret, i.e. which becomes optimal, at least in the long run, the limit (this is a modified version of "the learner's goal" as characterised by Shalev-Shwartz and Ben-David 2014, p.251):

Definition 3.6 (Relative Learnability)**.** Given some condition C, a hypothesis set $\mathcal{F}$ is online learnable (in a relative sense) iff there is an f_l such that f_l is not para-scientific/no clairvoyant (i.e. based on $\mathcal{Y}$) and for any prediction game G satisfying C with $\mathcal{F} \cup \{f_l\}$ and $\mathcal{Y}$ it holds for all $1 \leq i \leq n$:

$$\lim_{t\to\infty} succ_{l,t} - succ_{i,t} \geq 0$$

I.e.: f_l is (according to definition 2.17) access optimal in the long run in G.

Note that if a set of predictions or hypotheses $\mathcal{F}$ allows for absolute online learnability of $\mathcal{Y}$, then it allows also for relative online learnability, given the constraints for both notions are the same (condition C):

Corollary 3.7 (Absolute ⇒ Relative Learnability)**.** *Given one and the same condition C: If $\mathcal{F}$ allows for absolute online learnability of $\mathcal{Y}$, then it allows also for relative online learnability.*

Proof. If $\mathcal{F}$ allows for absolute online learnability, then there is a learning algorithm f_l such that $\lim_{t\to\infty} succ_{l,t} = 1$. Since 1 is the maximum, f_l cannot be outperformed by any other prediction method in the long run, i.e. $\lim_{t\to\infty} succ_{l,t} - succ_{i,t} \geq 0$. □

We have mentioned above that optimality or non-positive regret can be achieved by help of strict subadditive growth (with t) of the regret or even with sublinear growth (with t) of the regret. For this reason online learnability in the relative sense is sometimes characterised also by reference to the existence of an algorithm for the learner such that the learner's regret grows sublinearly with t (see Shalev-Shwartz and Ben-David 2014, p.251; and Rakhlin, Sridharan, and Tewari 2010, p.1). This is, however, a little bit stronger notion of *relative learnability*, since it does not allow for "valleys" in the learning process (in principle one can achieve access optimality in the long run, although, e.g. at the beginning, one's regret even grew superlinearly with t for some while). Learnability in the sense defined above is sometimes also called *Hannan consistency*, which demands in its strict form that the actual regret becomes negligible as t grows (see Cesa-Bianchi and Lugosi 2006, p.70). Since non-positive or negligible regret is sometimes also called 'no-regret', learnability in this sense is sometimes also characterised by the existence of a *no-regret algorithm* which is the so-called *no-regret property* (see Schapire 2012, p.169).

Note that the difference between absolute and relative learnability concerns mainly the learning target: Absolute learnability is concerned with learning the truth $\mathcal{Y}$, whereas relative learnability is concerned with learning the best hypotheses of $\mathcal{F}$, or even outperforming them. Although the learning target differs in both cases, the task for showing that a learning problem can be accomplished, i.e. that the truth $\mathcal{Y}$ or the best predictions or hypotheses $\mathcal{F}$ can be learned, consists in defining algorithms f_l which reach the truth or are optimal. This is what is investigated in online learning theory: "Our goal is to study which hypothesis classes are learnable in the online model, and in particular to find good learning algorithms for a given hypothesis class" (Shalev-Shwartz and Ben-David 2014, p.246). That, as we have mentioned above, the no-regret property is also called a 'consistency property', shows that no-regret algorithms play a fundamental role in online learning. Being a no-regret algorithm seems to be, so to say, the bare minimum an algorithm has to satisfy in order to be of interest for machine learners. Short run optimisation is the problem they are really after. As we will see soon, in epistemic engineering we can work already quite well with this *bare minimum*.

In the next section we will illustrate cases of online learnability in the absolute sense. Given that the hypothesis set contains the true hypothesis in a learning or prediction game G, one can quite easily verify that the truth ($\mathcal{Y}$ of G) is online learnable. We will then show that without this assumption, online classification fails with respect to relative online learnability (section 3.3). Finally, we will show that there is a powerful algorithm in online regression which allows for relative online learnability too (section 3.4).

3.2 A Gentle Start Towards an Optimality Result

Online learning gets quite fast quite complicated. For this reason we start with a simple example illustrating the main idea of the problem under consideration by discussing a relatively simple algorithm that aims at answering the question of learnability in a comprehensible way.

Recall, the two relevant ingredients of a prediction or learning game G with n predictors or hypotheses are the truth $\mathcal{Y}$, represented by values y_t with $t \in \mathbb{N}$, and the predictions or hypotheses $\mathcal{F}$, represented by values $f_{i,t}$ with $1 \leq i \leq n$. The question is whether we can define an algorithm which is not para-scientific and which figures out the true hypothesis, given the true method or hypothesis $\mathcal{Y}$ is in the set of prediction methods or hypotheses $\mathcal{F}$, i.e. given G is realisable? And the answer is: *Yes!* For reasons of simple illustration we assume for the remainder of this section that the prediction games under investigation are binary classification games with the values 0 and 1 (for the true outcomes as well as the predictors).

The idea of our first algorithm, the so-called *consistent algorithm*, is simple: At each round pick out the first hypothesis of the set of hypotheses which were always correct in past and predict in accordance with this method. Note that since this method is defined on the predictions of other methods, it is a meta-method. The definition of this first meta-method is as follows (after Shalev-Shwartz and Ben-David 2014, p.247):

Definition 3.8 (Consistency Algorithm)**.** Let G be a realisable classification game with the true values $\mathcal{Y}$ and the predictions or hypotheses $\mathcal{F}$. Furthermore, let $\mathcal{C}_t$ be recursively defined as the sequence of all predictors of $\mathcal{F}$ (ordered by their index) which were always correct until $t-1$. I.e:

- $\mathcal{C}_0 = \langle f_1, \ldots, f_n \rangle$
- $\mathcal{C}_t = \langle f_i : f_i \in \mathcal{C}_{t-1} \text{ and } f_{i,t-1} = y_{t-1} \rangle$ (ordered by index)

Then the *consistency algorithm* f_{cons} predicts the value that is predicted by the first method of $\mathcal{C}_t$:

$$f_{cons,t} = \underbrace{\mathcal{C}_{t_1}}_{\text{an } f}{}_{t}$$

For obvious reasons f_{cons} reaches the truth $\mathcal{Y}$, since its regret regarding the best, i.e. the true, hypothesis or method is bounded as follows:

Theorem 3.9 (Regret Bound for Consistency Algorithm)**.** *In the realisable case of a binary classification game G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ it holds for all $1 \leq i \leq n$:*

$$aregret_{\langle f_{cons}, f_i \rangle, t} \leq n - 1 \quad i.e. \quad succ_{f_{cons}, t} \geq 1 - \frac{n-1}{t}$$

Proof. Since in the realisable case $\mathcal{Y} \in \mathcal{F}$, there will be always at least one hypothesis or prediction method in the set of consistent hypotheses or prediction methods $\mathcal{C}$. Since we are considering classification games, we can count the number of mistakes m which f_{cons} makes until it reaches $\mathcal{Y}$ and imitates only true hypotheses or methods in $\mathcal{C}$ as follows: $m = \{t : f_{cons,t} \neq y_t\}$. Now, f_{cons} makes a mistake at t ($f_{cons,t} \neq y_t$) iff also the first method in $\mathcal{C}_t$ made a mistake at t. So, if we denote the set of imitated mistaken methods or hypotheses with $\mathcal{I}$, it holds: $|m| = |\mathcal{I}|$. In the worst case there is only one true hypothesis or method in $\mathcal{F}$ and its index is n. In this case $|\mathcal{I}| = n - 1$. Hence, in general $|m| \leq n - 1$. Note that in a binary classification game $|m| = \sum\limits_{t \in m} \ell_{f_{cons},t}$. Since in all $t \in \mathbb{N} \setminus m$ f_{cons} makes no mistakes (i.e. $\ell_{f_{cons},t} = 0$) we get $|m| = \sum\limits_{t \in \mathbb{N}} \ell_{f_{cons},t} \leq n - 1$. The best method in the setting f_i, at least the truth, has no loss: $\sum\limits_{t \in \mathbb{N}} \ell_{f_i,t} = 0$. Hence, by definition 2.15, $aregret_{\langle f_{cons}, f_i \rangle, t} \leq n - 1$. □

In interpreting the cumulative loss as the number of mistakes, we were making use of the so-called 0-1-*loss*, which is the most common loss measure in binary prediction and can be defined as follows:

Definition 3.10 (0-1 Loss)**.** ℓ is a 0-1 loss function iff

$$\ell(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

Throughout our consideration of online classification we will use this loss function.

The relation between the number of mistakes and the number of consistent hypotheses is described in table 3.1.

$\lvert m \rvert$	$=$	0	1	2	3	4	$\cdots$	n
$\lvert \mathcal{C} \rvert$	$\leq$	$n - 0$	$n - 1$	$n - 2$	$n - 3$	$n - 4$	$\cdots$	$n - n$

Table 3.1: Mistake bound of the consistency algorithm: Let $|m|$ be the number of mistakes of the algorithm, and let $|\mathcal{C}|$ be the number of consistent or errorless hypotheses or methods. At the beginning, before any prediction, $|\mathcal{C}| = n$ since $\mathcal{C}$ contains all n hypotheses or methods. With each mistake of the algorithm, also the imitated hypothesis or method is excluded from $\mathcal{C}$. In the realisable case $|\mathcal{C}| \geq 1$ ($\mathcal{Y} \in \mathcal{F}$ and $\mathcal{Y}$ is always errorless). Hence $n - |m| \geq 1$, hence also $|m| \leq n - 1$.

The regret of f_{cons} grows sublinearly with t which means that in the limit it vanishes. Hence, f_{cons} is a no-regret algorithm which means that in the realisable case $\mathcal{Y}$ is learnable (compare the general claims about online learnability of classes of hypotheses in Shalev-Shwartz and Ben-David 2014, pp.246ff):

Theorem 3.11 (Possibility of Absolute Learning)**.** *Given* realisability*, in a binary classification game G with the true values $\mathcal{Y}$ and a finite set of predictions $\mathcal{F}$ the true hypothesis or method $\mathcal{Y}$ is learnable (in the absolute sense).*

Proof. We simply spell out the argumentation of (Shalev-Shwartz and Ben-David 2014, pp.246ff): From the proof of theorem 3.9 we know that $\sum\limits_{t\in\mathbb{N}} \ell_{f_{cons},t} \leq n-1$. By arithmetic transformation we get for any $t \in \mathbb{N}$: $1-\ell_{f_{cons},1}+\cdots+1-\ell_{f_{cons},t} \geq -n+1+t$. Hence: $\frac{\sum\limits_{u=1}^{t} s_{f_{cons},u}}{t} \geq \frac{t+1-n}{t}$. Hence $succ_{f_{cons},t} \geq 1+\frac{1-n}{t}$. Hence $\lim\limits_{t\to\infty} succ_{f_{cons},t} = 1$. Hence, by definition 3.4: $\mathcal{Y}$ is online learnable (in the absolute sense). □

This is a very simple algorithm which allows for learning the true hypothesis, once it is in the hypothesis set (and the hypothesis set is finite). Considering its long run performance it is perfectly fine, since it is a no-regret algorithm. Regarding its short run performance it is not overwhelming, since, if n is high, its short run regret bound is also high. Can we do better? The answer is *Yes, we can!* How, will be shown in the next section. There we will also show that relaxing the condition of realisability leads to an impossibility result regarding relative online learnability within classification games.

3.3 Online Classification and Suboptimality

Note that throughout this section we make use of the 0-1 loss ℓ as characterised in definition 3.10:

$$\ell(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

Let us assume again that G is a binary classification game with the values 0 and 1, and that G is realisable, i.e. $\mathcal{Y} \in \mathcal{F}$. In the preceding section we have defined our first access optimal meta-method which allows also for absolute learnability: f_{cons}. Its guaranteed short run success is (see theorem 3.9)

$$succ_{f_{cons},t} \geq 1 - \frac{n-1}{t}$$

where $n = |\mathcal{F}|$ is the number of hypotheses or prediction methods of G. This means that only after $t = n-1$ rounds $succ_{f_{cons},t}$ is guaranteed to be positive, i.e. success is guaranteed. Now, we can do better by help of another algorithm, the so-called *halving algorithm*. The idea of *halving* is to not simply pick out one of the consistent prediction methods and predict accordingly. Rather, *halving* predicts in accordance with the simple majority of the up to now consistent hypotheses or prediction methods. The definition of f_{half} is as follows (see Shalev-Shwartz and Ben-David 2014, p.247):

Definition 3.12 (Halving Algorithm)**.** Let G be a classification game with the true values $\mathcal{Y}$ and the predictions or hypotheses $\mathcal{F}$. Furthermore, let $\mathcal{C}_t$ be recursively defined as the set of all predictors of $\mathcal{F}$ which were always correct until $t-1$. I.e:

- $\mathcal{C}_0 = \{f_1, \ldots, f_n\}$
- $\mathcal{C}_t = \{f_i : f_i \in \mathcal{C}_{t-1}$ and $f_{i,t-1} = y_{t-1}\}$

Then the *halving algorithm* f_{half} predicts the value which is predicted by the majority of $\mathcal{C}_t$:

$$f_{half,t} = \left\lceil \frac{\sum\limits_{f_i \in \mathcal{C}_t} f_{i,t}}{|\mathcal{C}_t|} \right\rfloor$$

Again, for obvious reasons f_{half} reaches the truth $\mathcal{Y}$ in the realisable case, since its regret regarding the best, i.e. the true, hypothesis or method is bounded as follows (see Shalev-Shwartz and Ben-David 2014, p.247):

Theorem 3.13 (Regret Bound for Halving Algorithm)**.** *In the realisable case of a binary classification game G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ it holds for all* $1 \leq i \leq n$*:*

$$aregret_{\langle f_{half}, f_i \rangle, t} \leq \lfloor \log_2(n) \rfloor \quad i.e. \quad succ_{f_{half},t} \geq 1 - \frac{\lfloor \log_2(n) \rfloor}{t}$$

Proof. (see Shalev-Shwartz and Ben-David 2014, p.247). Again, in the realisable case ($\mathcal{Y} \in \mathcal{F}$) $|\mathcal{C}| \geq 1$ (more specifically: for all t: $|\mathcal{C}_t| \geq 1$). Again, we can count the number of mistakes m which f_{half} makes until it reaches $\mathcal{Y}$ and imitates only true hypotheses or methods in $\mathcal{C}$: $m = \{t : f_{half,t} \neq y_t\}$. Now, f_{half} makes a mistake at t ($f_{half,t} \neq y_t$) iff also more than or 50% of methods in $\mathcal{C}_t$ made a mistake at t. So, at $t+1$ $\mathcal{C}_{t+1}$ will contain less than or 50% of the methods in $\mathcal{C}_t$: $|\mathcal{C}_{t+1}| \leq \frac{|\mathcal{C}_t|}{2}$. So, with each mistake of f_{half} the number of consistent hypotheses or methods $|\mathcal{C}|$ is at least halved. Since at the beginning $|\mathcal{C}_0| = n$, with $|m|$ mistakes $|\mathcal{C}_0|$ is at least halved $|m|$ times: $\frac{|\mathcal{C}_0|}{2^{|m|}}$ Halving will end in the worst case, when there is only one true hypothesis or method in $\mathcal{F}$ so at t when $|\mathcal{C}_t| = 1$. Hence $\frac{|\mathcal{C}_0|}{2^{|m|}} \geq 1$. Since $|\mathcal{C}_0| = n$, it follows that $|m| \leq \log_2(n)$. Note again that in a binary classification game $|m| = \sum\limits_{t \in m} \ell_{f_{half},t}$. Since in all $t \in \mathbb{N} \setminus m$ f_{half} makes no mistakes (i.e.

$\ell_{f_{half},t} = 0$) we get $|m| = \sum_{t\in\mathbb{N}} \ell_{f_{half},t} \leq \log_2(n)$. The best method in the setting f_i, at least the truth, has no loss: $\sum_{t\in\mathbb{N}} \ell_{f_i,t} = 0$. Hence, by definition 2.15, $\mathit{aregret}_{\langle f_{half}, f_i\rangle,t} \leq \log_2(n)$. □

The relation between the number of mistakes and the number of consistent hypotheses is described in table 3.2.

$\lvert m\rvert$	$=$	0	1	2	3	4	$\cdots$	n
$\lvert\mathcal{C}\rvert$	$\leq$	$n = \frac{n}{2^0}$	$\frac{n}{2} = \frac{n}{2^1}$	$\frac{\frac{n}{2}}{2} = \frac{n}{2^2}$	$\frac{\frac{\frac{n}{2}}{2}}{2} = \frac{n}{2^3}$	$\frac{\frac{\frac{\frac{n}{2}}{2}}{2}}{2} = \frac{n}{2^4}$	$\cdots$	$\frac{n}{2^n}$

Table 3.2: Mistake bound of the halving algorithm: Let $|m|$ be the number of mistakes of the algorithm, and $|\mathcal{C}|$ the number of consistent or errorless hypotheses or methods. At the beginning, before any prediction, $|\mathcal{C}| = n$ since $\mathcal{C}$ contains all n hypotheses or methods. With each mistake of the algorithm, also the imitated hypotheses or methods are excluded from $\mathcal{C}$. Since they were in the majority, with each mistake $\mathcal{C}$ is reduced by at least 50%, i.e. $|\mathcal{C}|$ reduces at least to its half. In the realisable case $|\mathcal{C}| \geq 1$ ($\mathcal{Y} \in \mathcal{F}$ and $\mathcal{Y}$ is always errorless). Hence $\frac{n}{2^{|m|}} \geq 1$, hence also $|m| \leq \log_2(n)$.

Clearly $\log_2(n) \leq n-1$ ($1.6, 3.5, 4.4, \ldots$ vs. $2, 10, 20, \ldots$), hence the halving algorithm allows for much better online learnability in the absolute sense, if the prediction game is realisable. So, we have a second meta-method which allows for online learnability in the absolute sense of any hypothesis set $\mathcal{F}$ with $|\mathcal{F}| = n$ such that $\lim_{n\to\infty} \log_2(n)$ exists. We will see soon that absolute online learnability is not restricted to finite hypothesis sets, so there are also some infinitely large hypotheses sets $\mathcal{F}$ which can be learned in an absolute sense. Furthermore, the notion of absolute online learnability is not trivial in the sense that there are hypotheses sets $\mathcal{F}$ which are not absolutely online learnable. This result will be employed later on when we proof an impossibility result regarding the traditional problem of epistemic justification.

Now, considering only the cardinality of $\mathcal{F}$, our halving algorithm is already optimal—also in the short run. The bound $\log_2(n)$ is the lowest regret bound which is guaranteed given we know only that $|\mathcal{F}| = n$. However, we can do better, once we not only consider the cardinality of $\mathcal{F}$, but also all of its structural information available to us. The idea is as follows: Consider the hypothesis set $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$ with $f_{i,t} = 1$ if $i = t$ and $f_{i,t} = 0$ otherwise. The sequences are depicted in the left part of table 3.3. Now, according to the bound we proved above, f_{half} makes at most $\log_2(4) = 2$ mistakes. However, in the realisable case an adversary (🐱) cannot fully employ this bound, since the structure of $\mathcal{F}$ is such that whenever the adversary makes f_{half} to err, also the majority, i.e. every hypothesis but one, makes an error and is ruled out by f_{half} as inconsistent. So, due to the specific structure of $\mathcal{F}$, an adversary can trick f_{half} at most once. Clearly, with a different hypothesis set $\mathcal{F}'$ with a different structure an adversary can fully employ the $\log_2(n)$-bound. Consider, e.g., the hypotheses $f'_1, \ldots, f'_4$ as defined in the right part of table 3.3: Here an adversary can, e.g., err f_{half} in the first round by setting $y_1 = 1 - f_{half,1} = 1 - 1 = 0$. In round 2 f_{half} predicts 1, so an adversary can simply err it by setting $y_2 = 1 - f_{half,2} = 1 - 1 = 0$, and still there is a hypothesis in the setting, namely f'_3, which did not make any error, i.e. the predication game is still realisable.

t	1	2	3	4	5	6	$\cdots$
$f_{1,t}$	1	0	0	0	0	0	$\cdots$
$f_{2,t}$	0	1	0	0	0	0	$\cdots$
$f_{3,t}$	0	0	1	0	0	0	$\cdots$
$f_{4,t}$	0	0	0	1	0	0	$\cdots$

t	1	2	3	4	5	6	$\cdots$
$f'_{1,t}$	1	1	0	0	0	0	$\cdots$
$f'_{2,t}$	0	1	1	0	0	0	$\cdots$
$f'_{3,t}$	0	0	1	1	0	0	$\cdots$
$f'_{4,t}$	1	0	0	1	0	0	$\cdots$

Table 3.3: Example of two hypothesis sets $\mathcal{F}$ and $\mathcal{F}'$. The former allows for better learnability via the halving algorithm due to specific structural features of $\mathcal{F}$: $\boldsymbol{aregret}_{\langle f_{half}, f_i \rangle, t} \leq 1$ (vs. $\log_2(4) = 2$). The latter does not allow for better learnability: $\boldsymbol{aregret}_{\langle f_{half}, f'_i \rangle, t} \leq 2 = \log_2(4)$

The exact structural feature relevant for determining an even more narrow bound is described in a, so to say, *logic of deceivability*: In the unrealisable case, an adversary (🐱) has any freedom to err the learning algorithm. In the realisable case, there is the restriction that the adversary cannot trick completely freely, but has to take care that there is always at least one hypothesis, the true hypothesis, which is without any error. Clearly, the realisable case is quite unrealistic; also from an epistemic point of view it allows only for discussing a quite weak form of scepticism. However, didactically speaking it is very valuable, because it allows one to gain easily insight into the *logic of deceivability*. This logic consists of combinatorial considerations regarding the hypothesis set $\mathcal{F}$ available to the learning algorithm. According to the upper bound provided in theorem 3.13, in the realisable case an adversary can err the learning algorithm f_{half} maximally $\log_2(|\mathcal{F}|)$ times. In order to approach this bound, we now allow the adversary not only to set the values y such that the learning algorithm f_l errs maximally (still satisfying the constraint of realisability), but we also allow her to prolongate the learning phase by freely picking the data in such a way that f_l learns the truth as late as possible. This makes a relevant difference. Consider, e.g., the left and the right sequences in table 3.4: In the left part, the learning algorithm receives already with the first event ($t = 1$) favourable information which allows it to easily learn the true hypothesis by just one single mistake—regardless at which round the adversary tries to trick it, after being tricked f_{half} figured out the true hypothesis: If the adversary were to trick f_{half} in round 1, i.e. $f_{half,1} = 0$, whereas $y_1 = 1$, then f_1, f_2, f_3 were inconsistent and ruled out. If at round 2, then f_4 is ruled out because of its inconsistency in round 1, and f_1, f_2 are ruled out after round 2 due to their inconsistency in this round. At round 3 the adversary cannot trick f_{half}, because f_4, f_3 would have been ruled out due to their inconsistency in round 1 and 2 respectively and f_1, f_2 (amongst them the true hypotheses) make the same prediction, so the setting would not be realisable. If it were at round 4, then f_2 is inconsistent at round 4 and only f_1 remained. And at round 5 the adversary can no longer trick f_{half} since f_1 dropped out due to its inconsistency in round 4 and the remaining hypothesis f_2 has to be the truth, due to the realisability condition. In the right part the first two events of the left part are just switched. By presenting the learner the second event first, the adversary can fully deploy the $\log_2(4) = 2$-bound and err f_{half} twice. In part II we will interpret this freedom of the adversary to

t	1	2	3	4	5
$f_{1,t}$	0	0	0	0	1
$f_{2,t}$	0	0	0	1	0
$f_{3,t}$	0	1	1	0	0
$f_{4,t}$	1	1	0	0	0
y_t	0	1	1	0	0
$f_{half,t}$	0	0	1	0	0

t'	1	2	3	4	5
$f_{1,t'}$	0	0	0	0	1
$f_{2,t'}$	0	0	0	1	0
$f_{3,t'}$	1	0	1	0	0
$f_{4,t'}$	1	1	0	0	0
y'_t	0	0	0	0	1
$f_{half,t'}$	1	0	0	1	1

Table 3.4: Example of the relevance of order for the possibilities of an adversary (🐱) to trick the learning algorithm f_{half}: In the left case, the order is favourable inasmuch f_{half} can find out the true hypothesis with just one single mistake—given this order of the series, the adversary can maximally trick f_{half} once. In the right case, the order is not favourable inasmuch as an adversary can trick the learning algorithm twice.

change the order of the events as a special form of Hume's claim that there is no logical connection between the past and the present.

How can we devise a logic of deceivability? Let us make some combinatorial considerations: Recall from theorem 3.13 that in a binary realisable classification game (realisability is a necessary condition for a prediction game to be relevant with regards to absolute online learnability) the upper bound for mistakes or regret of the best learning algorithm we know so far, f_{half}, is $\log_2(|\mathcal{F}|)$. In our analysis of table 3.4 above, we have already seen that the more balance there is between hypotheses predicting 0 and hypotheses predicting 1, the less hypotheses f_{half} can rule out once it is being tricked. If there is a complete balance in the remaining set of consistent hypotheses $\mathcal{C}$, i.e. if 50% of $f_i \in \mathcal{C}$ predict 0 and 50% of them predict 1, then only 50% are rule out afterwards. If the distribution is not balanced between these values at all, e.g., if all predict 0 or 1, then the adversary cannot trick the algorithm at all, since it needs to take care of the realisability assumption. These considerations allow for deriving two maxims for an adversary to trick the algorithm: The learning algorithm can be tricked at most $\log_2(|\mathcal{F}|)$ times. So, an adversary (🐱) should look out for finding $\log_2(|\mathcal{F}|)$ events in Y such that (although not explicitly mentioned, this seems to be the gist of the logic of deceivability as described in Shalev-Shwartz and Ben-David 2014, sect.21.1):

- The predictions of the hypotheses are maximally balanced between 0 and 1 in such a way that:
- They can be brought into an order such that the balance remains maximal regardless of cutting all hypotheses predicting 0 or all hypotheses predicting 1 at the round before.

Assume, e.g., that $n = |\mathcal{F}| = 8$, so $\mathcal{F}$ contains $f_1, \dots, f_8$ hypotheses. From theorem 3.13 we know that the adversary can trick f_{half} at most $\log_2(|\mathcal{F}|) = 3$ times. So, the adversary's task is to look out for 3 events of Y in accordance with the above maxims. Now, in a binary prediction game with 3 events and 8 hypotheses, best balancing clearly is achieved if there are 3 events such that all of the 8 hypotheses predict differently regarding the 3 events. In the worst case, there are no 3 events such that the hypotheses differ at all—so their predictions are all identical. The number of possible hypothesis sets regarding 3 events is calculated as the number of combinations of possibly different predictions with repetition: In the binary case, with respect to $n = 3$ events $k = 2^n = 8$ different predictions are possible. Since there are $n = 8$ hypotheses $f_1, \dots, f_8$ we can choose $n = 8$ elements from the set with $k = 8$ different predictions, where our choice is a *combination with possible repetition*. As we will see later on, in the general classificatory case of m possible predicted values, the maximal number of events where an adversary can err the learning algorithm is $\log_m(n)$, hence $k = m^{\log_m(n)} = n$ in general. So, the—for the adversary relevant—possible hypothesis space contains $\frac{(2 \cdot n - 1)!}{(n-1)! \cdot n!}$ elements, which are in the case of 8 hypotheses already 6435 combinations. Among them is the case where all 8 hypotheses predict differently, but also the case where all 8 hypotheses make the same predictions. In the former case the adversary can fully employ the $\log_2(n)$-mistake bound, in the latter she cannot err the learning algorithm at all. The example already shows that the considerations of the adversary are computationally speaking quite demanding. Later on we will devise a learning algorithm which incorporates the adversary's deceivability logic—and it is clear that also such an algorithm would be computationally quite demanding.

Now, our considerations from above are spelled out in the online learning literature by help of decision trees. For the following see mainly

(Shalev-Shwartz and Ben-David 2014, sect.21.1.1). The idea is as follows: The adversary (😈) decides among a decision tree. A binary decision tree consists of nodes, representing events, which are connected by edges, representing occurrence 1 or not-occurrence 0 of the events. Since in the binary case every event either occurs or does not occur, every node is connected via two edges to other nodes. The details of such a decision tree are presented in figure 3.1. We define the depth of such a tree as the number of edges in a path from the root to the leaf. As can be also seen in figure 3.1, a binary decision tree with depth n contains $2^{n+1} - 1$ nodes.

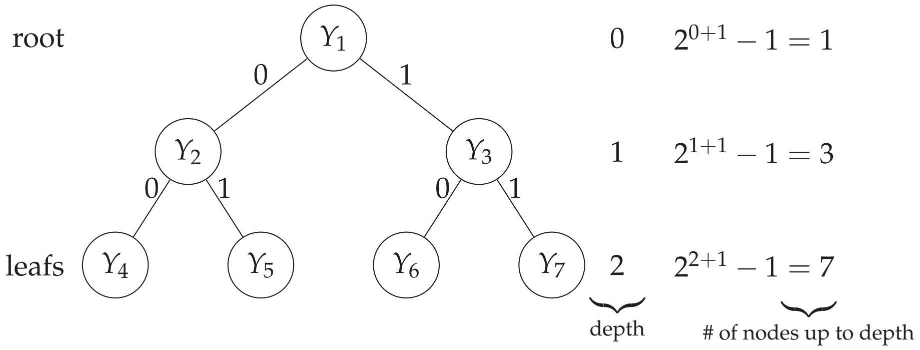

Figure 3.1: Example of a binary decision tree of depth 2 (number of edges in a path from the root to the leaf)

If we take the example from figure 3.1, in the binary case the adversary (😈) designs a prediction game as follows: She presents to the learning algorithm f_l and the other predictors some event Y_1 and asks for a prediction. She errs the learner by setting $y_1 = 1 - f_{l,1}$. If $y_1 = 0$, she goes on with the left child, i.e. Y_2, if $y_1 = 1$ she goes on with the right child, i.e. Y_3. In the first case, the adversary presents to the learner event Y_2 and asks for a prediction. Again, she errs the learner by setting $y_2 = 1 - f_{l,2}$. Then she goes on with Y_4 in case of $y_2 = 0$ and with Y_5 otherwise. Similarly in the second case of Y_3. At the end the events are relabelled such that they produce a sequence of events of a prediction game (e.g. $\langle 1, 2, 5 \rangle \mapsto \langle 1, 2, 3 \rangle$). As we have mentioned already above, the adversary might freely choose among the set of events Y and she can also permute the nodes of the decision tree.

Clearly, the adversary can err the learning algorithm in this way only, if she still satisfies the realisability condition, i.e. if she does not err all the other predictors. That she can freely err the learning algorithm without erring all the other predictors too, means that at each end node of such a binary decision tree (i.e. at each leaf) there are at least two hypotheses left that are consistent with their respective path: one stating that the leaf-event does not occur and one stating that it occurs. Every path that does not satisfy this condition also invalidates the realisability assumption. So, it is a path not viable for the adversary. This constraint for satisfying realisability is depicted in figure 3.2.

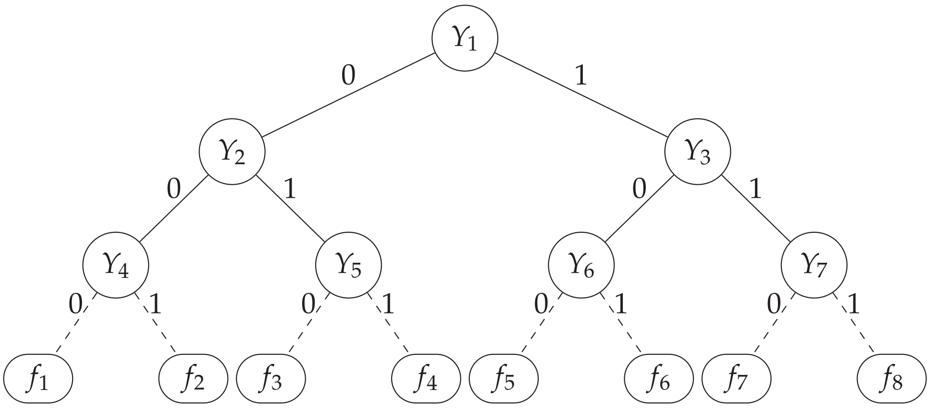

Figure 3.2: Binary decision tree with consistent hypotheses at each leaf, hence satisfying the realisability condition

Now we have all ingredients needed in order to make the task of the adversary 😈 explicit: An adversary has to look out for events such that she can order them in a decision tree in such a way that for each leaf of the tree at least two hypotheses remain, where one predicts 0 for the leaf-event and the other predicts 1. Furthermore, these hypothesis are consistent with their respective path. If this is the case, then it is said in the online learning literature that the prediction or hypothesis set $\mathcal{F}$ *shatters* the decision tree (see Shalev-Shwartz and Ben-David 2014, p.248). In order to employ her adversarial possibilities maximally, the task of the adversary is to find a *shattered* tree with maximal depth. This is the relevant structural feature of the hypothesis set $\mathcal{F}$ we were talking about above.

For illustrative purposes, let us consider an example: Consider the case with 8 hypotheses as discussed before. And let us assume that in the series of events we find $2^{\log_2(8)-1+1} - 1 = 8 - 1 = 7$ (we subtract from the log 1 because we started counting the depth of a tree with 0) events such that the 8 hypotheses make predictions as presented in table 3.5. Then this hypothesis set $\mathcal{F} = \{f_1, \ldots, f_8\}$ shatters the tree in figure 3.1 accordingly with figure 3.2.

Considering table 3.5, we see that there is some flexibility for the adversary in finding fitting events. However, this holds only at first glance. As the following theorem shows, the chance of an adversary to fully employ the $\log_2$-bound for erring a learning algorithm decreases drastically with the number of hypotheses:

Theorem 3.14 (Chances of Maximal Deceiving)**.** *In a realisable classification game with k possible values the chance for an adversary to be able to deceive the learning algorithm (f_{half},$f_{3\text{-}div}$ or $\cdots$ or $f_{k\text{-}div}$—see below) maximally according*

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
t	1	2	3	4	5	6	7
$f_{1,t}$	0	0	0|1	0	0|1	0|1	0|1
$f_{2,t}$	0	0	0|1	1	0|1	0|1	0|1
$f_{3,t}$	0	1	0|1	0|1	0	0|1	0|1
$f_{4,t}$	0	1	0|1	0|1	1	0|1	0|1
$f_{5,t}$	1	0|1	0	0|1	0|1	0	0|1
$f_{6,t}$	1	0|1	1	0|1	0|1	1	0|1
$f_{7,t}$	1	0|1	0	0|1	0|1	0|1	0
$f_{8,t}$	1	0|1	1	0|1	0|1	0|1	1

Table 3.5: Example of a hypothesis set shattering the decision tree as depicted in figure 3.1 accordingly with figure 3.2. $0|1$ means that for shattering the tree it does not matter whether the hypothesis predicts 0 or 1 for the respective event.

to the $\log_k(n)$ *regret bound is (for* $k \geq 2$*):*

$$\frac{1}{k^{n \cdot \lfloor \log_k(n) \rfloor}}$$

In the binary case with $k = 2$ *it holds:*

- *Given* $n = 1$ *hypothesis, the chance clearly is 0*
- *Given* $n = 2$ *hypotheses, the chances are 1/4*
- *Given* $n = 4$ *hypotheses, the chances are 1/256*

Proof. Regarding the generality of k ($k > 2$) see theorem 3.25 below. Regarding the binary case we demonstrate the theorem by help of the example from above: In the realisable case, a hypothesis class with $|\mathcal{F}| = n = 8$ hypotheses allows for maximally $\log_2(n) = 3$ mistakes. In order to fully employ this mistake bound, the adversary needs to provide $n - 1 = 7$ events such that at each of the $2^{\log_2(n)-1} = n/2 = 4$ leafs of a decision tree with these 7 events is covered by two hypotheses, consistent with the full path to the root. So we need to employ all of the $2 \cdot n/2 = n = 8$ hypotheses. Since there are 7 events in the tree, this allows for $2^{n-1} = 2^7 = 128$ different possible prediction series. Since we have 8 hypotheses, we need to combine 8 such series, i.e. regarding the 7 events we have $\left(2^{n-1}\right)^n = 128^8$ possible hypothesis sets. As we easily verify by considering table 3.5 or figure 3.2, each of the 8 hypothesis of $\mathcal{F}$ is relevant only for predicting $\log_2(n) = 3$ events. The remaining events remain underdetermined with respect to the shattering property. Hence, for each hypothesis $n - 1 - \log_2(n) = 7 - 3 = 4$ predictions can be freely varied. Since there are $n = 8$ hypotheses we get $n \cdot (n - 1 - \log_2(n)) = 8 \cdot 4 = 32$ parameters which can be varied between

0 and 1 without any harm for the shattering property. Hence, given $n = 8$ hypotheses, out of the $\left(2^{n-1}\right)^n = 128^8$ possible combinations of prediction series for $n - 1 = 7$ events, there are only $2^{n\cdot(n-1-\log_2(n))} = 2^{32}$ combinations of prediction series which allow for shattering a binary decision tree with the 7 events. So, the chances are:

$$\frac{k^{n\cdot(n-1-\log_k(n))}}{\left(k^{n-1}\right)^n} = \frac{1}{k^{n\log_k(n)}}$$

□

Now, as theorem 3.14 shows, the $\log_2$-bound is more of theoretical than practical relevance to an adversary (🐱). Furthermore, the consideration of shattered decision trees sheds also some light on this bound—so to say, it provides an illustrative explanation of this bound and shows that it is the "limiting case" of a decision tree: Given $|\mathcal{F}| = n$ hypotheses, an adversary can construct in the best case a shattered decision tree where all end nodes, i.e. the leafs, are covered by at least two hypotheses. Hence, such a shattered decision tree has at most $n/2$ leafs. Now, in a binary decision tree, exactly two leafs share one parent node, hence, the number of parent nodes of the leafs is $(n/2)/2 = n/2^2$. Again, exactly two parent nodes of leafs share one parent node (a grand parent of a leaf), so the number of grand parents of a leaf is $((n/2)/2)/2 = n/2^3$ and so on, until we reach the overall ancestor, namely the root node, which is 1 element. Hence, the depth d of a tree, i.e. the number of edges on a path from a leaf to the root, is given by $n/2^d = 1$. Resolving for d we get $d = \log_2(n)$. Since an adversary can only trick a learning algorithm along one path, and the number of edges of a path, i.e. the depth of a tree, represents the number of mistakes the learning algorithm makes (see figures 3.1 and 3.2) we get as bound for the mistakes $\log_2(n)$.

Now, given this result about the role of the structure of the hypothesis class $\mathcal{F}$, we can also provide narrower bounds: First, assume that an adversary errs a learning algorithm d times. So the depth of the binary decision tree is d. As described in figure 3.1 such a tree contains $2^{d+1} - 1$ events of Y. Let us relabel them from top to bottom and left to right by $Y_1, \ldots, Y_{2^{d+1}-1}$. Now, such a tree is shattered, if for every path there are hypothesis $f_i, f_j \in \mathcal{F}$ which are consistent with the path and cover the leafs of the path. Now, as can be seen in figure 3.1, paths are specific sequences of events: $\langle Y_1, Y_2, Y_4\rangle$ is a path, and also $\langle Y_1, Y_2, Y_5\rangle$, but not, e.g., $\langle Y_1, Y_2, Y_3\rangle$. Due to our labelling, whether a sequence of events is a path or not depends on the value we assign to an event: In assigning $y_1 = y_2 = 0$, we characterise $\langle Y_1, Y_2, Y_4\rangle$ as a path, similarly $y_1 = y_3 = 1$ characterises $\langle Y_1, Y_3, Y_7\rangle$ as path. Given our labelling, the recursive definition of a path depending on the true values y on it can be defined by the following index function (taken from Shalev-Shwartz and Ben-David 2014, p.249):

Definition 3.15 (Index Function for Path Generation)**.**

$$index(t) = 2^{t-1} + \sum_{u=1}^{t-1} y_u \cdot 2^{t-1-u}$$

We verify this by considering examples:

- Let $t = 1$: Then $index(t) = 2^0 = 1$, so:
  - $Y_{index(t)}$ refers to event Y_1
- Let $t = 2$: Then $index(t) = 2^1 + y_1 \cdot 2^0 = 2 + y_1$, so:
  - If $y_1 = 0$: $Y_{index(t)}$ refers to event Y_2
  - If $y_1 = 1$: $Y_{index(t)}$ refers to event Y_3
- Let $t = 3$: Then $index(t) = 2^2 + y_1 \cdot 2^1 + y_2 \cdot 2^0 = 4 + 2y_1 + y_2$, so:
  - If $y_1 = 0, y_2 = 0$: $Y_{index(t)}$ refers to event Y_4
  - If $y_1 = 0, y_2 = 1$: $Y_{index(t)}$ refers to event Y_5
  - If $y_1 = 1, y_2 = 0$: $Y_{index(t)}$ refers to event Y_6
  - If $y_1 = 1, y_2 = 1$: $Y_{index(t)}$ refers to event Y_7
- And so forth . . .

So, *index* encodes a binary decision tree. We can use the *index*-function not only for referring to events Y, but also to the predictions of the hypotheses or methods in $\mathcal{F}$ (we suppose that these were relabelled with the relabelling of the events). So if, e.g., $t = 2$ and $y_1 = y_2 = 0$, then $f_{i,index(t)}$ is the predicted value of f_i for event Y_4, i.e. f_i's prediction for round 4. Now, given this codification of decision trees, we can define the notion of a *shattered tree* precisely (modification of Shalev-Shwartz and Ben-David 2014, pp.248f):

Definition 3.16 (Shattered Decision Tree)**.** A sequence $T = \langle Y_1, \ldots, Y_{2^{d+1}-1} \rangle$ is a binary decision tree of depth d that is shattered by $\mathcal{F}$ iff for any $\langle y_1, \ldots, y_d \rangle \in \{0,1\}^d$ there are $f_i, f_j \in \mathcal{F}$ such that for all $t \in \{1, \ldots, d\}$ it holds: $f_{i,index(t)} = f_{j,index(t)} = y_t$ and $f_{i,index(t+1)} \neq f_{j,index(t+1)}$.

Note that the binary decision tree of depth 2 in figure 3.2 is shattered by $\mathcal{F} = \{f_1, \ldots, f_8\}$ as described in table 3.5, because for all $\langle y_1, y_2 \rangle \in \{0,1\}^2$ there are two consistent hypotheses covering the leafs—details are presented in table 3.6.

Whether a set of hypotheses $\mathcal{F}$ shatters a given tree or not is easily verified: We have to verify whether each leaf is covered by at least two—regarding the leaf contradicting—hypotheses? If we also want to know the

y_1	y_2	Consistent Covering Hypotheses
0	0	f_1, f_2
0	1	f_3, f_4
1	0	f_5, f_6
1	1	f_7, f_8

Table 3.6: Shattering according to definition 3.16. E.g.: In case $y_1 = 0$ and $y_2 = 0$, $f_{1,index(1)} = f_{1,1}$, $f_{1,index(2)} = f_{1,2}$, $f_{1,index(3)} = f_{1,4}$. Analogously: $f_{2,index(1)} = f_{2,1}$, $f_{2,index(2)} = f_{2,2}$, $f_{2,index(3)} = f_{2,4}$. According to the definition of table 3.5: $f_{1,1} = f_{1,2} = f_{1,4} = 0$ and $f_{2,1} = f_{2,2} = 0$ whereas $f_{2,4} = 1$. Hence, f_1, f_2 satisfy the condition in definition 3.16 for the case $y_1 = 0 = y_2$. Similarly for the other cases and the other hypotheses.

sub-trees of a given unshattered tree which are shattered by $\mathcal{F}$ or a subclass of it, we just need to remove stepwise levels—leaf by leaf so to say—until we end up with leafs each of which have contradicting hypotheses. The set of all these hypotheses shatters the remaining sub-tree—see, e.g., figure 3.3.

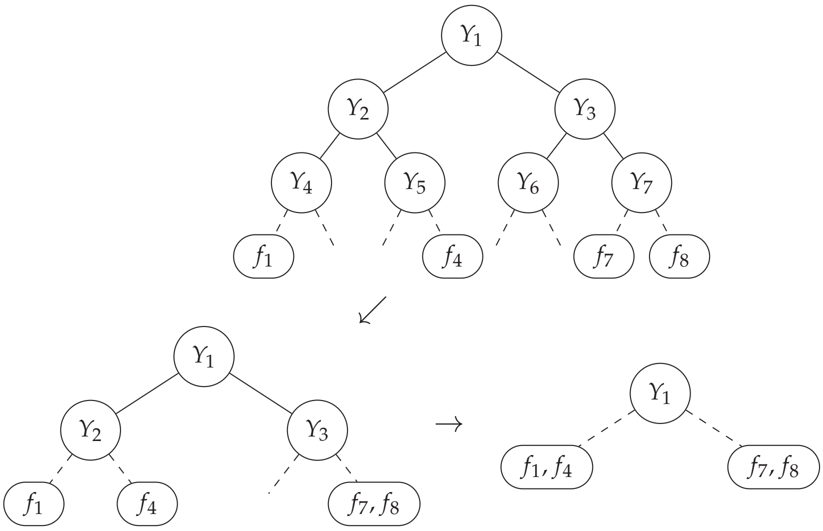

Figure 3.3: Verifying the shattering property of a hypothesis set $\mathcal{F}$ with reduction to subtrees: Given $\mathcal{F} = \{f_1, \dots, f_8\}$ we might verify that $\mathcal{F}$ does not shatter the upper tree from above. So, we remove the leaf-level and verify whether the hypotheses allow for shattering the remaining tree (depicted on the left side). We can iterate this step until we reach at least the root (depicted on the right side). In this example $\{f_1, f_7\}$, $\{f_1, f_8\}$, $\{f_4, f_7\}$, $\{f_4, f_8\}$, and any superset thereof shatters the tree containing only Y_1 as root and leaf node.

Clearly, every sub-tree of a shattered decision tree is shattered too by one and the same hypothesis set. So, e.g., the subtree $Y_4 \leftarrow Y_2 \rightarrow Y_5$ of the decision tree in figure 3.2 is also shattered by $\{f_1, f_3\}$, $\{f_1, f_4\}$, $\{f_2, f_3\}$, $\{f_2, f_4\}$, and any superset thereof as, e.g., $\mathcal{F}$. An adversary is not only interested in finding such a sub-tree, but to fully exploit the potential of $\mathcal{F}$. For this reason, her task is it not to find any by $\mathcal{F}$ shattered binary decision tree, but such a tree with maximal depth. For this purpose we define a quantitative measure for the relevant structural property of a hypothesis set, namely the shattering property. Since this measure goes back to computer scientist Nick Littlestone, it is also called *Littlestone's dimension* (Littlestone himself called it an "optimal mistake bound" see Littlestone 1988; Shalev-Shwartz and Ben-David 2014, p.249):

Definition 3.17 (Littlestone's Dimension Ldim)**.** Let $\mathcal{F}$ be the prediction or hypothesis set of a prediction game G. Then $Ldim(\mathcal{F})$ is the maximal integer d such that there exists a binary decision tree T consisting of a sequence of events of G which has depth d and is shattered by $\mathcal{F}$. (If there is no such tree, $d = 0$.)

As we have argued above, the $\log_2$-bound is the "limiting case" of shattering. This means that the $\log_2$-bound is the upper bound of Ldim (see Shalev-Shwartz and Ben-David 2014, p.249):

Corollary 3.18 (Upper Bound of Ldim)**.** *For any $\mathcal{F}$ of some prediction game G: $Ldim(\mathcal{F}) \leq \lfloor \log_2(|\mathcal{F}|) \rfloor$*

From definitions 3.16 and 3.17 it follows that $Ldim(\mathcal{F})$ is also the minimal value of any upper bound for the regret of a learning algorithm (see Lemma 21.6 in Shalev-Shwartz and Ben-David 2014, p.249):

Theorem 3.19 (Ldim As Lower Bound)**.** *Given a prediction or hypothesis set $\mathcal{F}$ of a prediction game G, any upper bound for the regret of a learning algorithm as, e.g., f_{cons} (with an upper bound of $|\mathcal{F}| - 1$) or f_{half} (with an upper bound of $\lfloor \log_2(|\mathcal{F}|) \rfloor$) is $\geq Ldim(\mathcal{F})$.*

Proof. (Sketch due to restricted theoretical framework) The idea behind this theorem is that binary decision trees cover all possibilities of deceiving and the shattering property covers the guarantee. Let G be a realisable prediction game with the true hypothesis $f_i \in \mathcal{F}$ of G and let d be the guaranteed number after which a learning algorithm f_l makes no longer a mistake, i.e. d is the maximal regret of f_l compared to f_i. This means that the proof of the upper bound rules out any possibility of erring f_l after d. Now, all possibilities of erring f_l are captured by decision trees that are shattered by $\mathcal{F}$ (this covers all possible combinations and without shattering no guarantee is possible). Now, assume $d < Ldim(\mathcal{F})$. $Ldim(\mathcal{F})$ is defined as the maximal integer d' such that there is a binary decision tree T of depth d' such that $\mathcal{F}$ shatters T ($Ldim(\mathcal{F}) = d'$). This means that f_l can be erred also until $d' > d$, contradicting the claim that d is the guaranteed upper bound. Hence, $d \geq Ldim(\mathcal{F})$. □

Whenever all hypotheses f_i, f_j of $\mathcal{F}$ predict equivalently, i.e. for all t: $f_{i,t} = f_{j,t}$, then $\mathcal{F}$ allows for no shattering, hence $Ldim(\mathcal{F}) = 0$. The limiting case is where $|\mathcal{F}| = 1$, where the realisability constraint avoids deceiving. For illustrative purposes, consider also the following examples of (Shalev-Shwartz and Ben-David 2014, p.249):

- Let $\mathcal{F} = \{f_1, \ldots, f_n\}$ be such that $f_1, \ldots, f_n$ cover all possible predictions regarding $\log_2(n)$ events (thinking in terms of truth tables this means that each row of such a table is occupied by one of $f_1, \ldots, f_n$). Then $Ldim(\mathcal{F}) = \log_2(|\mathcal{F}|)$.
- Let $\mathcal{F} = \{f_1, \ldots, f_n\}$ such that $f_i = 1$, if $i = t$ and $f_i = 0$, if $i \neq t$. I.e.: Every hypothesis predicts 1 at exactly one point in time and no two hypotheses predict 1 at the same time. Then $Ldim(\mathcal{F}) = 1$: At each t where no f_i predicts 1, the adversary cannot err the learner. And at each round t where exactly one f_i predicts 1, the learner just needs to predict 0. If the adversary errs the learner, then the learner figured out the true hypothesis. If the adversary does not err the learner, then the learner excluded one wrong hypothesis without making an error.
- Furthermore, let $\mathcal{F} = \{\{\langle i, t, x\rangle : t \in \mathbb{N} \;\&\; x = 1, \text{if } i = t \;\&\; x = 0, \text{if } i \neq t\} : i \in \mathbb{N}\}$. Then $Ldim(\mathcal{F}) = 1$ due to the same reason as above. Note, however, that here $\mathcal{F}$ is not finite and hence $\lim_{n \to |\mathcal{F}|} \log_2(n) = \infty$.

Recall our explanation of the $\log_2$-bound from above: Given a prediction game G with $\mathcal{F}$ that is realisable (assume f_i to be the true predictor or hypothesis), we are considering a binary decision tree of depth d which is shattered by $\mathcal{F}$. We saw that shattering implies that there are two hypotheses at each leaf of the tree, hence a shattered tree contains at most $|\mathcal{F}|/2$ leafs. Going back from the leafs to parent nodes, grandparent nodes etc. is equivalent to repeatedly halving the number of nodes. We end up at the root (one single node) after d-times iterated halving, i.e. $|\mathcal{F}|/2^d = 1$. Resolving for d brought about the $\log_2$-bound. Note that this does not only illustrate the bound, but also the learning algorithm for which we have proven the bound, namely the halving algorithm f_{half}: By at least halving the number of correct methods $\mathcal{C}$ (with initially $\mathcal{C} = \mathcal{F}$) iteratively with every step it has being erred, f_{half} approached the root f_i after at most d steps, where $|\mathcal{C}|/2^d = 1$.

Now, similarly as described for the halving case, we can devise a learning algorithm for approaching the $Ldim$-bound: We know that $Ldim(\mathcal{F})$ expresses the maximal number of erring possibilities for an adversary. So, we, as learners, aim at reducing these possibilities. Given two hypothesis classes $\mathcal{F}_1$ and $\mathcal{F}_2$ we prefer that one with the lower Littlestone's dimension $Ldim$, because this means we can be erred less often. However, this implies that in making a prediction we opt for that one of the hypothesis class with higher *Ldim*: If the adversary errs me in making this choice, she has ruled out also the unfavoured hypothesis class. And if she does not err me, I am still with my unfavoured choice, but got rid of the other class of hypotheses, and this completely for free (without making a mistake). So, according to this idea one should not just pick the majority, but that hypothesis set with higher *Ldim*. The definition of the so-called *standard optimal algorithm* f_{soa} is as follows (see Littlestone 1988; Shalev-Shwartz and Ben-David 2014, p.250):

Definition 3.20 (Standard Optimal Algorithm)**.** Let G be a classification game with the sequence of true values $\mathcal{Y}$ and the sequences of predictions $\mathcal{F}$. Furthermore, let $\mathcal{C}_t$ be recursively defined as the set of all predictors of $\mathcal{F}$ which were always correct until $t-1$. I.e:

- $\mathcal{C}_0 = \{f_1, \ldots, f_n\}$
- $\mathcal{C}_t = \{f_i : f_i \in \mathcal{C}_{t-1} \text{ and } f_{i,t-1} = y_{t-1}\}$

Furthermore, based on $\mathcal{C}$ define $\mathcal{C}^0$ and $\mathcal{C}^1$ as:

- $\mathcal{C}_t^0 = \{f_i : f_{i,t} = 0\} \cap \mathcal{C}_t$
- $\mathcal{C}_t^1 = \{f_i : f_{i,t} = 1\} \cap \mathcal{C}_t$

Then:

$$f_{soa,t} = \begin{cases} 1 & \text{if } Ldim(\mathcal{C}_t^1) \geq Ldim(\mathcal{C}_t^0) \\ 0 & \text{otherwise} \end{cases}$$

This algorithm has the following regret bound and for this reason allows for absolute online learnability (see Lemma 21.7 Shalev-Shwartz and Ben-David 2014, p.250):

Theorem 3.21 (Regret Bound for Standard Optimal Algorithm)**.** *In the realisable case of a binary classification game G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ it holds for all $1 \leq i \leq n$:*

$$aregret_{\langle f_{soa}, f_i \rangle, t} \leq Ldim(\mathcal{F}) \quad i.e. \quad succ_{f_{soa},t} \geq 1 - \frac{Ldim(\mathcal{F})}{t}$$

Proof. We show that whenever f_{soa} is erred at t, then also *Ldim* of the set of consistent hypotheses is reduced at least by 1. I.e.: $Ldim(\mathcal{C}_{t+1}) \leq Ldim(\mathcal{C}_t) - 1$ in case $f_{soa,t} \neq y_t$. We can count the number of mistakes m which f_{soa} makes until it reaches $\mathcal{Y}$ and imitates only true hypotheses or methods in $\mathcal{C}$: $m = \{t : f_{soa,t} \neq y_t\}$. Given that *Ldim* of the next chosen hypothesis set is at least reduced by 1 if f_{soa} is erred, this means that whenever m grows by 1, *Ldim* of the next chosen hypothesis set is reduced. Since the minimal *Ldim* is 0, m can grow maximally to *Ldim* of the hypothesis set $\mathcal{F}$: $|m| \leq Ldim(\mathcal{F})$.

Now, let us show indirectly that $Ldim(\mathcal{C}_{t+1}) \leq Ldim(\mathcal{C}_t) - 1$, by assuming $Ldim(\mathcal{C}_{t+1}) > Ldim(\mathcal{C}_t) - 1$, although f_{soa} was erred at t, i.e. $f_{soa,t} \neq y_t$. Clearly, *Ldim* cannot grow with a constant or decreasing hypothesis set, and since by definition 3.20 $\mathcal{C}_{t+1} \subseteq \mathcal{C}_t$, we get $Ldim(\mathcal{C}_{t+1}) = Ldim(\mathcal{C}_t)$. But then also $\mathcal{C}_{t+1} = \mathcal{C}_t$. However, since f_{soa} was erred at t, at least some competing fs must have been erred too for which reason $\mathcal{C}_{t+1} \subset \mathcal{C}_t$. Hence, $Ldim(\mathcal{C}_{t+1}) \neq Ldim(\mathcal{C}_t)$.

In a binary classification game $|m| = \sum\limits_{t \in m} \ell_{f_{soa},t}$. Since in all $t \in \mathbb{N} \setminus m$ f_{soa} makes no mistakes (i.e. $\ell_{f_{soa},t} = 0$) we get $|m| = \sum\limits_{t \in \mathbb{N}} \ell_{f_{soa},t} \leq Ldim(\mathcal{F})$. The best method in the setting f_i, at least the truth, has no loss: $\sum\limits_{t \in \mathbb{N}} \ell_{f_i,t} = 0$. Hence, by definition 2.15, $\boldsymbol{aregret}_{\langle f_{soa}, f_i \rangle, t} \leq Ldim(\mathcal{F})$. □

For a comparison of f_{half} and f_{soa}, consider table 3.7, with an example for the suboptimality of the halving algorithm in comparison to the standard optimal algorithm.

	Y_1	Y_2	Y_3	Y_u	Y_{u+1}	Y_{u+2}
t	1	2	3	u	$u+1$	$u+2$
$f_{1,t}$	1	0	0	1	0	0
$f_{2,t}$	0	1	1	0	1	1
$f_{3,t}$	1	0	0	1	0	0
$f_{4,t}$	0	0	1	0	0	1
y_t	0	0	1	0	0	1
$f_{half,t}$	1	1	1	0	0	1
$f_{soa,t}$	0	1	1	0	0	1

Table 3.7: Example of the suboptimality of f_{half} compared to f_{soa}: The prediction game consists of $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$ and $\mathcal{Y}$ with y as defined above (for any $3 < u \in \mathbb{N}$). The game is realisable, since there is a true hypothesis in $\mathcal{F}$, namely f_4. The halving algorithm simply takes the prediction of the majority (and in tie cases 1) of consistent hypotheses, which means that in this example it errs 2 times (that is also the maximum $\log_2(4)$) until it learns the true hypothesis f_4. The standard optimal algorithm, on the other hand, chooses the consistent hypothesis set with maximal *Ldim* (and in tie cases 1). In round 1, the consistent hypothesis classes are $C_1^1 = \{f_1, f_3\}$ and $C_1^0 = \{f_2, f_4\}$. Since $f_{1,t} = f_{3,t}$, $Ldim(C_1^1) = 0$, whereas $Ldim(C_1^0) = 1$. By predicting accordingly with C_1^0, f_{soa} makes only one error until it learns the true hypothesis f_4. The errors are marked grey.

Note that f_{soa} is a meta-method which is computationally very demanding. It is not a meta-inductive method, since it makes use also of information about future predictions. So to say, f_{soa} must have access to its competitors' prediction methods in the sense of knowing their algorithms or all past, present, and future predictions. However, this does not imply that f_{soa} is para-scientific, since access to the truth is not at all needed for calculating *Ldim*. f_{soa} needs access only to the true values of past events when devising the set of consistent hypotheses. By combining theorems 3.19 and 3.21, we get the following result about the absolute optimality of f_{soa} in the realisable case:

Corollary 3.22 (Realisable Classification Bound)**.** *Given a prediction or hypothesis set $\mathcal{F}$ of a realisable prediction game G, the best achievable guaranteed upper and lower bound for an (not para-scientific) online learning algorithm is that of f_{soa} whose guaranteed upper and lower bound for regret is $Ldim(\mathcal{F})$.*

Furthermore, we can even characterise the notion of *absolute online learnability* by help of *Ldim* (see Shalev-Shwartz and Ben-David 2014, pp.250f):

Theorem 3.23 (Characterisation of Absolute Online Learnability)**.** *The conditions (C) for a hypothesis set $\mathcal{F}$ of a prediction game G to allow for absolute online learnability are exactly those:*

(CA1) G is realisable, and:

(CA2) $Ldim(\mathcal{F})$ is finite.

Proof. Regarding realisability (CA1) we have argued already that otherwise there were no constraint preventing the adversary to simply define the truth as $y_t = 1 - f_{l,t}$. So, (CA1) is necessary for absolute learnability. Regarding finiteness of *Ldim* (CA2) it holds: By definition 3.4 we know that $\mathcal{F}$ allows for online learnability iff there is a not para-scientific learning algorithm f_l such that in any realisable prediction game with hypothesis $\mathcal{F}$ $\lim_{t\to\infty} succ_{l,t} = 1$. Now, by corollary 3.22 we get for any such f_l and any t in the realisable case: $succ_{l,t} \leq succ_{soa,t}$. Hence, $\lim_{t\to\infty} succ_{l,t} = 1$, only if $\lim_{t\to\infty} succ_{soa,t} = 1$. Now, by theorem 3.21 it holds for any t: $succ_{soa,t} \geq 1 - \frac{Ldim(\mathcal{F})}{t}$. Hence, $\lim_{t\to\infty} succ_{soa,t} = 1$ only if $Ldim(\mathcal{F})$ is finite. Hence, also (CA2) is a necessary condition of absolute learnability. That both are sufficient for absolute learnability follows from theorem 3.21. □

Given the combinational considerations from above, we can generalise the results for the case of classification. Up to now we were mainly concerned with so-called *binary classification*. In the following we will indicate how to expand these results to so-called *multiclass classification*, i.e. classification with more than two values: Here the relevant decision trees are not

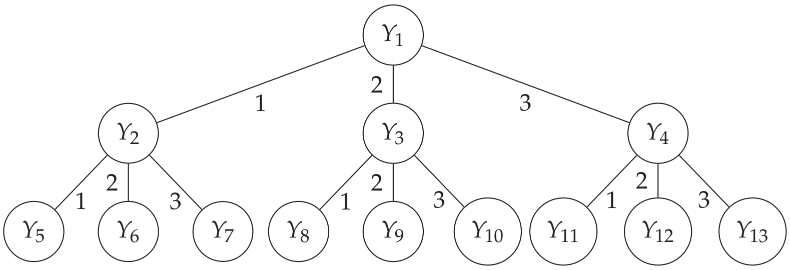

Figure 3.4: Example of a ternary decision tree of depth 2 (number of edges in a path from the root to the leaf)

binary, but—given k classes—k-ary. An example of a ternary decision tree is provided in figure 3.4.

Similarly as above, for every classification with k values that can be predicted, we can define an algorithm $f_{k\text{-}div}$ which predicts accordingly with the majority:

Definition 3.24 (Divide-by-k Algorithm)**.** Let G be a classification game with the true values $\mathcal{Y}$ $(v_1, \ldots, v_k)$ and the predictors or hypotheses $\mathcal{F}$. Furthermore, let $\mathcal{C}_t$ be recursively defined as the set of all predictors of $\mathcal{F}$ which were always correct until $t-1$. I.e:

- $\mathcal{C}_0 = \{f_1, \ldots, f_n\}$
- $\mathcal{C}_t = \{f_i : f_i \in \mathcal{C}_{t-1}$ and $f_{i,t-1} = y_{t-1}\}$

Then the *k-div algorithm* $f_{k\text{-}div}$ predicts the value which is predicted by the majority of $\mathcal{C}_t$:

$$f_{k\text{-}div,t} = \begin{cases} v_1 & \text{iff } |\{f_i : f_i \in \mathcal{C}_t \,\&\, f_{i,t} = v_1\}| > \\ & \quad |\{f_i : f_i \in \mathcal{C}_t \,\&\, f_{i,t} = r\}| \\ & \quad \text{for all } v_1 \neq r \in \{v_1, \ldots, v_k\} \\ & \qquad\qquad \vdots \\ v_{k-1} & \text{iff } |\{f_i : f_i \in \mathcal{C}_t \,\&\, f_{i,t} = v_{k-1}\}| > \\ & \quad |\{f_i : f_i \in \mathcal{C}_t \,\&\, f_{i,t} = r\}| \\ & \quad \text{for all } v_{k-1} \neq r \in \{v_1, \ldots, v_k\} \\ v_k & \text{otherwise} \end{cases}$$

This general algorithm enjoys the following regret bound (this is a generalisation of the regret bound for realisable binary classification games as provided in Shalev-Shwartz and Ben-David 2014, p.247):

Theorem 3.25 (Regret Bound for Classification in General)**.** *In the realisable case of a k-ary classification game G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ it holds for all $1 \leq i \leq n$:*

$$\mathit{aregret}_{\langle f_{k\text{-}div}, f_i \rangle, t} \leq \lfloor \log_k(n) \rfloor \quad \textit{i.e.} \quad \mathit{succ}_{f_{k\text{-}div}, t} \geq 1 - \frac{\lfloor \log_k(n) \rfloor}{t}$$

Proof. The proof is analogous to the one for f_{half}: In the realisable case ($\mathcal{Y} \in \mathcal{F}$) $|\mathcal{C}| \geq 1$ (more specifically: for all t: $|\mathcal{C}_t| \geq 1$). We can count the number of mistakes m which $f_{k\text{-}div}$ makes until it reaches $\mathcal{Y}$ and imitates only true hypotheses or methods in $\mathcal{C}$: $m = \{t : f_{k\text{-}div,t} \neq y_t\}$. Now, $f_{k\text{-}div}$ makes a mistake at t ($f_{k\text{-}div,t} \neq y_t$) iff also more than or $100/k\%$ of methods in $\mathcal{C}_t$ made a mistake at t. So, at $t+1$ $\mathcal{C}_{t+1}$ will contain less than or $100/k\%$ of the methods in $\mathcal{C}_t$: $|\mathcal{C}_{t+1}| \leq \frac{|\mathcal{C}_t|}{k}$. So, with each mistake of $f_{k\text{-}div}$ the number of consistent hypotheses or methods $|\mathcal{C}|$ is at least divided by k. Since at the beginning $|\mathcal{C}_0| = n$, with $|m|$ mistakes $|\mathcal{C}_0|$ is at least divided by k $|m|$ times: $\frac{|\mathcal{C}_0|}{k^{|m|}}$. Dividing by k will end in the worst case, when there is only one true hypothesis or method in $\mathcal{F}$ so at t when $|\mathcal{C}_t| = 1$. Hence $\frac{|\mathcal{C}_0|}{k^{|m|}} \geq 1$. Since $|\mathcal{C}_0| = n$, it follows that $|m| \leq \log_k(n)$. Note again that in a binary classification game $|m| = \sum\limits_{t \in m} \ell_{f_{k\text{-}div},t}$. Since in all $t \in \mathbb{N} \setminus m$ $f_{k\text{-}div}$ makes no mistakes (i.e. $\ell_{f_{k\text{-}div},t} = 0$) we get $|m| = \sum\limits_{t \in \mathbb{N}} \ell_{f_{k\text{-}div},t} \leq \log_k(n)$. The best method in the setting f_i, at least the truth, has no loss: $\sum\limits_{t \in \mathbb{N}} \ell_{f_i,t} = 0$. Hence, by definition 2.15, $\mathit{aregret}_{\langle f_{k\text{-}div}, f_i \rangle, t} \leq \log_k(n)$. □

Note that the higher k, the better the performance of the algorithm: In the binary case ($k = 2$) we ruled out of two hypotheses f_1, f_2 with different predictions just one (1 out of 2). In the 3-ary case we rule out of three hypotheses f_1, f_2, f_3 with different predictions already two (2 out of 3). And more generally, in the k-ary case, with high enough k, almost all (also: all but one) differing hypotheses are ruled out. More generally, this fact is expressed by theorem 3.14 about the chances of maximal deceiving which decrease drastically with k.

Now, also for the case of multiclass classification one can define the notion of *shattering a k-ary tree* straightforward analogously to definition 3.16. Only the *index*-function has to be modified and the covering-condition: Each leaf has to be covered by k hypotheses predicting k different values for the leaf (but are still consistent with one and the same path). And given this notion of shattering a k-ary tree a quantitative measure *Ldim* can be defined straightforward analogously to definition 3.17. Then one can devise a learning algorithm f_{soa} based on this *Ldim* analogously to definition 3.20.

Et Voilà! One ends up with the provably best *guaranteed* learner for classification in general. And hence also a perfect characterisation of absolute online learnability by the finiteness of *Ldim* accordingly to theorem 3.23.

Note that these results indicate already that absolute learnability can be achieved also in the case of online regression: Very sloppily speaking, if we consider online regression as a limiting case where the number of possible values k approaches infinity (we might represent this—although not well defined—by $\lim\limits_{k\infty} \log_k(|\mathcal{F}|)$), then regret also decreases and approaches 1, i.e. an adversary can maximally err the learner once.

Up to now we were concerned with absolute online learnability only. Around our discussion of corollary 3.5 we have stated that realisability is a necessary condition for absolute learnability (otherwise a daemon could trivially err the learner by setting $y_t = 1 - f_{l,t}$). In the characterisation result of theorem 3.23 we have stated another necessary condition, namely finiteness of *Ldim*. So these are the two conditions which are necessary and sufficient for absolute online learnability. Now, let us come to the notion of relative online learnability? Is relative online learnability in the classification case possible? Recall, relative online learnability means that there is a no-regret learning algorithm, i.e. an algorithm whose regret vanishes in the limit (or becomes negative), or equivalently: an algorithm which approaches or even overshoots the success rates of the best prediction methods or hypotheses in any prediction game. Regarding realisable prediction games any algorithm that allows for online learnability in the absolute sense also allows for online learnability in the relative sense, since in the realisable case the algorithm reaches the true hypotheses and by this gains maximal success in the long run (see corollary 3.7). So, regarding relative learnability we need to concentrate on unrealisable prediction games.

Unrealisable prediction games are those whose hypothesis sets do not contain the true hypothesis. Since from a strict epistemic stance it does not matter whether the truth is *not* in our hypothesis class or whether we *do not know* whether it is in or not, in online learning such prediction games are also called *agnostic* (see Shalev-Shwartz and Ben-David 2014, p.245).

There is a very nice parallel between the realisable case and absolute online learnability on the one hand, and the case of assuming a best expert in the setting and relative learnability on the other: If one assumes that there is a best prediction method or hypothesis in $\mathcal{F}$, then we can simply restate the task for the adversary. Recall, in case of absolute learning the task for an adversary was to try to err a learning algorithm f_l as often as realisability allows. The *logic of deceivability* (i.e. the logic applied by 🐱) as described above uncovers that an adversary's possibilities are restricted to shattered trees and *Ldim*. In case of relative learning the restated task for an adversary is to try to err a learning algorithm f_l as often as the *existence of a best expert*-constraint allows. As we briefly discuss now, the underlying *logic of deceivability* remains the same. Let us first begin with a definition of the games we are interested in:

Definition 3.26 (Best Expert Game)**.** A prediction game G with $\mathcal{Y}$ and $\mathcal{F}$ is a best expert game iff there is exactly one $f_i \in \mathcal{F}$ such that for $j \in \{1, \ldots, n\} \setminus \{i\}$ and all $t \in \mathbb{N}$:

$$succ_{i,t} \geq succ_{j,t}$$

Since the logic of deceivability from above was only about the strict realisable case, we also define here only a strict notion of a best expert game. Clearly, every realisable prediction game is also a best expert game:

Corollary 3.27 (Realisable ⇒ Best Expert)**.** *If a prediction game G is realisable, then G is also a best expert game.*

When we introduced learning algorithms for absolute learnability in the realisable case, we aimed at devising an algorithm f_l which learns the true hypothesis and by this reaches $\lim_{t \to \infty} succ_{l,t} = 1$. Now, looking for learning algorithms for relative learnability we aim at devising an algorithm f_l which learns the best expert hypothesis and by this reaches $\lim_{t \to \infty} succ_{l,t} = \lim_{t \to \infty} succ_{b,t}$, where the best expert is f_b. If the best expert hypothesis is the truth—as, e.g., is the case in realisable games in accordance with corollary 3.27—, then $\lim_{t \to \infty} succ_{l,t} = \lim_{t \to \infty} succ_{b,t} = 1$. Now, we can mimic the role realisability plays for absolute learnability by the best expert assumption and relative learnability as follows: In the case of absolute learnability we included the truth $\mathcal{Y}$ into the hypothesis set $\mathcal{F}$, e.g. by defining an $f_i \in \mathcal{F}$ as $f_{i,t} = y_t$. In the case of relative learnability we simply reverse the direction and replace, so to say, the truth by the best expert hypothesis f_b: We define $y_t = f_{b,t}$. Note that we assumed that in the best expert case there is exactly one best expert. Otherwise the choice of an expert for defining the true series would not be unique and an adversary (🐱) could enter the scenery again. The task of the learning algorithm f_l is to learn this "truth" absolutely, i.e. $\lim_{t \to \infty} succ_{l,t} = 1$. This embeds *best expert classification* with relative online learnability into *realisable classification* with absolute online learnability. The underlying logic is the same, also the algorithms, hence, we can provide a partial characterisation of the notion of *relative online learnability* by help of the following conditions (this is the gist of Shalev-Shwartz and Ben-David 2014, sect.21.1.1):

Theorem 3.28 (Partial Characterisation of Relative Online Learnability)**.** *The following conditions (C) allow for relative online learnability of a hypothesis set $\mathcal{F}$ of a prediction game G:*

(CR1) G is a best expert game, and:

(CR2) $Ldim(\mathcal{F})$ is finite.

Proof. By help of the above embedding we can use any algorithm devised for absolute learning also for relative learning in a best expert game. In the embedding $\mathcal{Y}$ is defined on basis of f_b, the predictions of the expert hypothesis. Conditions for absolute learnability of so-defined $\mathcal{Y}$ are realisability—which holds by definition of $\mathcal{Y}$—and finiteness of Littlestone's dimension of the hypothesis set: $Ldim(\mathcal{F})$. By this we end up with the conditions stated in theorem 3.28. Taken together they are sufficient for relative online learnability of $\mathcal{F}$. □

Now, can we account for relative online learnability also without these conditions? As we will briefly show now, relaxing condition (CR1) further by allowing for any prediction games is not possible in the *online classificatory* setting (note, as stated in corollary 3.27, (CR1) is already a weakening of the realisability condition (CA1)). In the following section on *online regression* we will see that there is a regression algorithm which allows for relative online learnability in any prediction game whatsoever.

An impossibility result for relative learnability in agnostic non-expert online classification can be traced back to the work of Thomas Cover (see Cover 1965; the historical claim comes from Shalev-Shwartz and Ben-David 2014, p.252). The example is as follows: Consider such a prediction game with $\mathcal{F} = \{f_1, f_2\}$ where these methods are defined as constant hypotheses: For all $t \in \mathbb{N}$: $f_{1,t} = 1$ and $f_{2,t} = 0$. Now, for any learning algorithm f_l which aims at relatively learning $\mathcal{F}$, define $\mathcal{Y}$ via the adversarial strategy: $y_t = 1 - f_{l,t}$. Note, since the prediction game can be agnostic, the adversary need not take care of realisability, etc. Clearly, f_l never scores and hence has no predictive success at all. On the other hand, at least one of f_1, f_2 will always score from time to time. Hence, f_l is no no-regret algorithm:

Theorem 3.29 (Impossibility of No-Regret Classification)**.** *There is* no *no-regret learning algorithm f_l which is not para-scientific.*

Proof. Assume f_l is a learning algorithm and not para-scientific. Take the example from above with $\mathcal{F} = \{f_1, f_2\}$. For any round t, we can calculate the regret as follows (see Shalev-Shwartz and Ben-David 2014, p.252): By definition of f_1, f_2, at each round either f_1 or f_2 scores. Define $t^+ = \sum\limits_{u=1}^{t} y_t$. If $t^+ \geq t/2$ then f_1 scored $\geq t/2$ times. Otherwise f_2 scored $\geq t/2$ times. Hence, the best hypothesis f_i (or hypotheses) have a cumulative loss $\leq t/2$ ($\sum\limits_{u=1}^{t} \ell_{i,u} \leq t/2$). Also by definition of $\mathcal{Y}$, f_l scores never, hence $\sum\limits_{u=1}^{t} \ell_{l,u} = t$. So, at round t the regret of f_l with regard to f_1 and f_2 is $\geq t - t/2 = t/2$.

Since $(t/2)/t = t^2/2$ is superlinear, regret of f_l grows with each round: $\lim_{t\to\infty} \boldsymbol{aregret}_{\langle l,i\rangle,t} = \lim_{t\to\infty} t^2/2 = \infty$. This example can be also extended to the case of $k > 2$-ary prediction games. □

Before we move on to the next section, one note regarding the regret bounds and the loss function is in place: First of all, we only operated with the 0-1 loss as defined in definition 3.10. One can use a different loss function as, e.g., a .25-1 loss function which does not penalise wrong predictions fully. One might interpret such a loss as providing some incentive for predicting at all. This can be relevant, e.g., for so-called *intermittent* prediction games where the prediction series do not have to be complete (some predictors might sometimes abstain from predicting). This is a highly relevant topic which is not covered by our investigation (for an investigation of this topic see, e.g., Schurz 2019, sect.7.2). However, it is important to note that the bounds for learning (absolute, not relative) with such loss functions that do not penalise wrong predictions fully become better: The learning algorithm's consistency classes need not be varied—in the realisable case the learner is still after the truth and hence excludes wrong hypotheses strictly. However, in case of the .25-1 loss every fourth mistake of the learner is, so to say, free (compared to the 0-1 loss). Hence, regret with respect to the true hypothesis is to a higher degree sublinear.

Regarding the guaranteed bounds we want to highlight that these bounds as, e.g., the *Ldim* bound or the $\log_2$ bound do not show that any algorithm is guaranteed to perform within these bounds. On the contrary, as we have seen, f_{half} sometimes exceeds the *Ldim* bound and f_{cons} almost regularly will exceed the $\log_2$ bound. These results hold only for the respective algorithms. Regarding the *Ldim* bound: We have seen that this is an upper bound as well as a lower bound for regret (or vice versa for success). That it is a lower bound for regret means that no algorithm is *guaranteed* to have less regret. And that it is an upper bound for regret means that there is an algorithm which is *guaranteed* to have at most so much regret. 'Guarantee' means that this holds for all prediction games (satisfying the respective condition C). So, regarding the *Ldim* upper bound this means that there is an algorithm such that for all realisable prediction games with $\mathcal{F}$ the algorithm regrets $\leq Ldim(\mathcal{F})$ times not having made a different choice ($\exists f_l \forall G$). The standard optimal algorithm f_{soa} was such an algorithm. And regarding the *Ldim* lower bound this means that for any learning algorithm there is a prediction game with $\mathcal{F}$ such that the algorithm regrets $\geq Ldim(\mathcal{F})$ times not having made a different choice ($\forall f_l \exists G$). Figuratively speaking, these bounds are about the best learners fighting the best adversaries. An upper bound concerns the best learner's part: "Fighting against" the whole spectrum of worst (most simple minded) and the best adversaries, the best learner's "costs" will always be below or equal to the bound. And a lower bound concerns the best adversary's part: "Fighting against" the spectrum of worst and best learners, the best adversary's (🐱) "harming" will always be above or equal to the bound.

Let us briefly recap the intermediate results of the logic of deceivability. Regarding online classification we have:

- If G with $\mathcal{F}$ is a realisable classification game, then $\mathcal{F}$ is absolutely and relatively online learnable, as long as $Ldim(\mathcal{F})$ is finite.
- If G with $\mathcal{F}$ is a best expert classification game, then $\mathcal{F}$ is *not* absolutely, but relatively online learnable, as long as $Ldim(\mathcal{F})$ is finite.
- If G with $\mathcal{F}$ is no best expert game (i.e. also not realisable), but an agnostic classification game, then $\mathcal{F}$ is *neither* absolutely, *nor* relatively online learnable.

Now, this is only a very rough scheme of optimisation and learnability in online classification. Since we are mainly after online regression, we do not aim at drawing a complete map of this field. However, we want to provide at least a short overview of more specific results of the theory of meta-induction for the case of online classification:

- Schurz (2008b, thrm.1) shows that *simple* meta-inductive learning in the sense of "imitating the best candidate method" (here the learner just copies the prediction of the up to now best expert—see Schurz 2019, sect.6.1, thrm.6.1) also allows for relative learnability; this can be guaranteed also for games where a best expert shows up only later on at some point in time t (this is more general than our characterisation of an expert classification game).
- The result from above can be even more generalised to prediction games where, starting with some point in time t, there is not only one best expert, but also a set of ε-best experts whose success rates deviate from each other only by at most ε. Schurz (2008b, thrm.2) and (Schurz 2019, sect.6.2, thrm.6.3) show that there is an ε-cautious learning algorithm—switching from an old best expert to a new best expert only, if the difference between the success rates of both passes the ε-threshold—which allows for (ε-)approximation of relative learnability.
- Finally, of its kind completely novel and even more general is the result about learning based on "deception recording" as proven in Schurz (2008b, thrm.3). The idea is that the learner not only keeps track of the success rates, but also of the success rates conditional on cases where the candidate method was the learner's favourite. If a candidate method has significantly more (unconditional) success than conditional success, i.e. if the candidate method is much more

often successful in cases where it is not imitated by the learner than in cases where it is imitated, then the learner labels it as a deceiver and ignores its predictions until the two success rates are balanced better. Such an "imitate the best non-deceiver" learning algorithm is proven to be access-optimal compared to all non-deceiving methods. The moral seems to be: In online classification there is no general route to optimality since there always might be deceivers around. But at least there is a remedy for optimality with regards to non-deceivers: namely imitating the best non-deceivers (for details on this see Schurz 2019, sect.6.3, thrm.6.6).

- Very interesting is also the result of Cesa-Bianchi and Lugosi (sect.4.3 2006) who show how learning by "following the perturbed leader" allows for ruling out systematic deceiving via implementing perturbations (which are independent from the true outcome/unknown to the deceiver beforehand; we will say a bit more on optimality in online classification under such an assumption in section 4.1);

When we discussed the traditional epistemic approaches to the problem of justification, we mentioned that an epistemic engineer's approach consists of seeking optimal means for some given ends (section 1.4). And if there are no optimal means, then one might try to cautiously and conservatively redefine the ends. So, if absolute or relative learnability are our ends, do we need to redefine them given the above impossibility? As we will argue in the next section and the subsequent chapter, regarding relative learnability the answer is: *No*! Up to now we have not exploit our framework fully (and for a machine learner it seems that we even have not scratched it).

3.4 Online Regression and Optimality

Recall, online regression consists in providing predictions within the interval $[0, 1]$. We have defined the notion of an online regression game and that of an online classification game in such a way that both are disjoint. However, since $0, 1 \in [0, 1]$, online classification can be also considered as a special case of online regression. The relevant difference is that in online classification the methods or hypothesis (including the learner) are allowed to provide predictions with only k different values, whereas in online regression there is no such restriction except that the values of the predictions are within the unit interval. This implies that for any number k of predicted different values, the learner can always predict a further different value $(k + 1)$. This allows for more flexibility on the learner's part, and so one might wonder whether this brings further restrictions for an adversary with it? However, note that also the adversary gains more flexibility, and it seems even more, inasmuch as she can choose a further different value ($k + 2$) and declares it as the true outcome. So, at first glance this seems to even strengthen the *logic of deceivability*. But that is only *deceptive*: As we will see in this section, the increased flexibility of the learner outperforms that of the adversary.

We have seen that in online classification there is hope for absolute learnability only, if the prediction game is a realisable one. And that there is hope for relative learnability, if the prediction game is a best expert game. Otherwise an adversary (🐱) can construct a Cover-style prediction game in which the learner always errs and where at least one prediction method or hypothesis has success in the long run to at least some degree. In case of binary classification with $\mathcal{F} = \{f_1, f_2\}$ we assumed that for all $t \in \mathbb{N}$ $f_{1,t} = 1$ and $f_{2,t} = 0$ and defined $y_t = 1 - f_{l,t}$, where f_l is the learning algorithm. Regardless of how one defines f_l, the learner is guaranteed not to score at all. The learning algorithm is guaranteed also to be not optimal. But even worse, if f_l is not defined as a function predicting equally to f_1 or f_2 in the long run, it is even guaranteed to be strongly suboptimal in the sense that it is guaranteed to be outperformed by f_1 as well as f_2. Now, if we allow for online regression, things are different. If we define f_l as just taking the average of the predictions of f_1 and f_2, then we achieve already a better performance. Let us define for all $t \in \mathbb{N}$: $f_{l,t} = (f_{1,t} + f_{2,t})/2$. Furthermore, let us assume the so-called *natural loss* ($\ell(x, y) = |x - y|$). Then, clearly, f_l will also score to some degree. Furthermore, f_l is no longer strongly suboptimal, since its predictions are always within that of f_1, f_2. It is, however, also not guaranteed to be optimal. This is due to its simple definition—note that f_l is a very simple meta-method which is still not inductive since it does not take into account any information about past outcomes and predictions. Nevertheless, what is relevant to note is that here the increased flexibility of the learning algorithm to predict a new value increased also its performance at least a little bit—from strong suboptimality to simple suboptimality. And note that the adversary could not at all employ her increased flexibility, because any different definition of the true outcome than simply taking a maximum distance from f_l's prediction would have brought an even better score for f_l while at the same time it would have decreased the score of the better hypothesis.

It is a general strategy in online regression to mix predictions in such a way that the adversary's (🐱) hands/paws are tied. And it is the possibility to mix which relevantly strengthens a learner's options in online regression compared to online classification. However, note also from the example above that once we allow for a significant increase of predictable values so that *real* (valued) mixing is possible, also the question of how to score becomes more relevant. And this means that our choice of the loss function is crucial. The best mixing strategy counts for nothing, if we still penalise any deviation from the true outcome fully (0-1 loss). Recall, in the example above we used the *natural loss*. If we had used the 0-1 loss, then f_l's mixing would be of no use to avoid strong suboptimality.

Now, there is a family of loss functions which one way or another encode that mixing is advantageous. It is the family of *convex* loss functions. A convex loss function is defined as follows:

Definition 3.30 (Convex Loss)**.** ℓ is a loss function which is convex (in its first argument) iff for all $x_1, \ldots, x_n$, $w_1, \ldots, w_n$ such that $\sum\limits_{i=1}^{n} w_i = 1$ and $w_i \geq 0$, and for all y it holds:

$$\ell\left(\sum_{i=1}^{n}(w_i \cdot x_i), y\right) \quad \leq \quad \sum_{i=1}^{n} w_i \cdot \ell(x_i, y)$$

I.e.: The loss of a weighted average of predictions is smaller than or equal to the weighted average of the losses of the predictions.

Throughout this section we will make use only of convex loss functions ℓ. As is easy to see, the 0-1 loss is not convex: If, e.g., for two predictions the weights $w_1, w_2 = 1 - w_1$ are not extreme, i.e. if $0 < w_1 < 1$, then the loss of two weighted opposed predictions will be always greater than the weighted loss of the opposed predictions:

$$\underbrace{\ell(w_1 \cdot 1 + w_2 \cdot 0, y)}_{=1 \text{ given a 0-1 loss}} \quad > \quad \underbrace{w_1 \cdot \ell(1, y) + w_2 \cdot \ell(0, y)}_{=w_1<1 \text{ or } =w_2<1 \text{ given a 0-1 loss}}$$

On the other hand, the natural loss is convex since for any weights $w_1, \ldots, w_n$ as specified in definition 3.30 it holds:

$$\left|\sum_{i=1}^{n}(w_i \cdot x_i) - y\right| \quad = \quad \sum_{i=1}^{n} w_i \cdot |x_i - y|$$

Now, let us come back to the example from above: the averaging learning algorithm. We have seen that, given two constant predictors and the natural loss function, this learning algorithm is at least not strongly suboptimal in the sense that it is not outperformed by all other predictors. Does this hold only for the chosen example? As we show now, the answer is: *No*! Any convex loss function excludes strong suboptimality of the averaging learner. In this sense the convexity of a loss function encodes that mixing is advantageous: It prevents strong suboptimality for the averaging algorithm. The averaging algorithm can be defined as follows:

Definition 3.31 (Averaging Algorithm)**.** Let G be a regression game with the true values $\mathcal{Y}$ and the set of predictions or hypotheses $\mathcal{F}$ (with $|\mathcal{F}| = n$). Then the averaging learning algorithm f_{av} is defined as:

$$f_{av,t} = \frac{\sum\limits_{i=1}^{n} f_{i,t}}{n}$$

Note that f_{av} is again a meta-method. It is, as mentioned above, not inductive. Still, it generally excludes being tricked by an adversary in the form of being strongly suboptimal:

Theorem 3.32 (No Strong Suboptimality of Averaging)**.** *Given a convex loss function ℓ and a prediction game G with a hypothesis set $\mathcal{F}$ ($|\mathcal{F}| = n$) it holds: f_{av} is* not *strongly suboptimal in the sense that for all $t \in \mathbb{N}$ there is some $1 \leq i \leq n$:*

$$\boldsymbol{aregret}_{\langle av,i\rangle,t} \leq 0$$

Proof. (Indirectly) Assume that ℓ is convex and that for all $1 \leq i \leq n$ and some t:

$$\sum_{u=1}^{t} \ell_{i,u} < \sum_{u=1}^{t} \ell_{av,t}$$

Then:

$$\sum_{i=1}^{n}\sum_{u=1}^{t} \ell_{i,u} < n \cdot \sum_{u=1}^{t} \ell_{av,u} \text{ hence } \sum_{u=1}^{t}\sum_{i=1}^{n} \ell_{i,u} < n \cdot \sum_{u=1}^{t} \ell_{av,u}$$

Hence:

$$\sum_{u=1}^{t}\sum_{i=1}^{n} \frac{1}{n} \cdot \ell_{i,u} < \sum_{u=1}^{t} \ell_{av,u} \text{ hence } \sum_{i=1}^{n} \frac{1}{n} \cdot \ell_{i,v} < \ell_{av,v}$$

Now, by definition 3.31:

$$\ell_{av,v} = \ell\left(\sum_{i=1}^{n} \frac{1}{n} \cdot f_{i,v}, y_v\right) \text{ hence } \sum_{i=1}^{n} \frac{1}{n} \cdot \ell_{i,v} < \ell\left(\sum_{i=1}^{n} \frac{1}{n} \cdot f_{i,v}, y_v\right)$$

But this contradicts the convexity assumption about ℓ, according to which:

$$\ell\left(\sum_{i=1}^{n} \frac{1}{n} \cdot f_{i,v}, y_v\right) \leq \sum_{i=1}^{n} \frac{1}{n} \cdot \ell_{i,v}$$

Hence, there is some $1 \leq i \leq n$ such that:

$$\sum_{u=1}^{t} \ell_{i,u} \geq \sum_{u=1}^{t} \ell_{av,t} \text{ hence } \boldsymbol{aregret}_{\langle av,i\rangle,t} \leq 0$$

□

From this follows immediately the following possibility result of avoiding strong suboptimality in the online regression setting:

Theorem 3.33 (Possibility of Avoiding Strict Subotimal Regression)**.** *Given the loss function is convex, there is a learning algorithm f_l, e.g. the averaging algorithm f_{av}, which allows for avoiding strong suboptimal regression.*

We have seen that in agnostic online classification every learning algorithm can easily become strongly suboptimal. Here we see that mixing in online regression allows for tying an adversary's hand such that she fails already with respect to a quite simple learner as f_{av} when it comes to strong suboptimality. However, although avoiding strong suboptimality is a reasonable desideratum for epistemic justification, it is clearly not enough for arguing by help of optimality. Can we do better than that and provide an algorithm which does not only avoid suboptimality, but even guarantees optimality? Yes, we can: The idea is as follows: The averaging algorithm presented above mixed the predictions in a constant way by assigning each prediction method a constant weight $1/n$. Now, in order to do better, we need to provide more sophisticated weights. Since we are after optimality, the idea is to make the weights success-dependent which is the same as making them regret dependent. The more relative success a prediction method had in past, the higher its weight. Which is to say that the higher the regret, the higher the weights. For technical convenience and in order to provide better bounds it is common practice to design weights in such a way that the regrets are in the exponent. As we will show now, such a learning algorithm allows for no-regret learnability, i.e. relative online learnability. This makes it not only impossible for an adversary to trick the learner such that she is only strongly suboptimal. It even ties the adversary's (🐱) hands such that she cannot avoid long run access optimality of the learner. Here are the details of the exponentially weighting learning algorithm:

Definition 3.34 (Exponential Weighting)**.** We define c, the exponential (e) cumulative (recursion with $\cdot$) loss (ℓ) of a prediction method or hypothesis (i) which is used for learning (η), recursively as follows (for all $1 \leq i \leq n$, $t \in \mathbb{N}$):

- $c_{i,1} = 1$
- $c_{i,t+1} = c_{i,t} \cdot e^{-\eta \cdot \ell_{i,t}}$

Based on the exponential cumulative loss we define weights by normalisation:

$$w_{i,t} = \begin{cases} \frac{c_{i,t}}{\sum\limits_{j=1}^{n} c_{j,t}} & \text{if the denominator} > 0 \\ \frac{1}{n} & \text{otherwise} \end{cases}$$

Given these weights we define the exponential meta-inductive forecaster as:

$$f_{emi,t} = \sum_{i=1}^{n} w_{i,t} \cdot f_{i,t}$$

Now, let us come to the main result applied in our investigation: The exponential meta-inductive learner f_{emi} is a no-regret algorithm. We prove this in two steps: First, we show that its regret up to some given point in time (or horizon) T grows only sublinearly with T. Then we show that this result can be generalised for any point in time which allows us to consider its regret in the limit. Let us start with the first step (see Shalev-Shwartz and Ben-David 2014, p.253):

Theorem 3.35 (Regret Bound^T for Exponential Weighting Algorithm)**.** *Given the learning parameter $\eta = \sqrt{\frac{2 \cdot \ln(n)}{T}}$ and a loss function ℓ which is convex in its first argument, it holds for all prediction games G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ ($|\mathcal{F}| = n$): For all $1 \leq i \leq n$:*

$$\mathit{aregret}_{\langle emi,i \rangle,T} \leq \sqrt{2 \cdot \ln(n) \cdot T}$$

Proof. Shalev-Shwartz and Ben-David (2014, pp.253f) provide a proof for the upper regret bound of the weighted majority algorithm in online classification, where a learning algorithm is allowed to make probabilistic predictions (randomised across the possible values). In section 4.1 we will describe this approach to overcome suboptimality in online classification. Since the probabilities are within $[0, 1]$, this is technically the same as providing a prediction within online regression. For this reason we can simply apply their proof also for theorem 3.35. Here is a slight modification of their proof:

1. Let $\eta = \sqrt{\frac{2 \cdot \ln(n)}{T}}$. Furthermore let ℓ be convex.
2. By definition of c in definition 3.34 we get the following equalities about the ratio of the denominators used in normalisation (the normalising denominator for $t+1$ and that of t):

$$\frac{\sum\limits_{i=1}^{n} c_{i,t+1}}{\sum\limits_{i=1}^{n} c_{j,t}} = \sum_{i=1}^{n} \frac{c_{i,t+1}}{\sum\limits_{j=1}^{n} c_{i,t}} = \sum_{i=1}^{n} \frac{c_{i,t} \cdot e^{-\eta \cdot \ell_{i,t}}}{\sum\limits_{j=1}^{n} c_{i,t}}$$

$$= \sum_{i=1}^{n} w_{i,t} \cdot e^{-\eta \cdot \ell_{i,t}}$$

3. By the inequality $e^{-x} \leq 1 - x + \frac{x^2}{2}$ (valid for all $x \geq 0$) we get the instance:
$$e^{-\eta \cdot \ell_{i,t}} \quad \leq \quad 1 - \eta \cdot \ell_{i,t} + \frac{\eta^2 \cdot \ell_{i,t}^2}{2}$$
Note that due to the assumptions in 1. $0 \leq \eta < 1$ and due to the boundedness of loss ℓ by $[0,1]$ (axiom 2.2) $\eta \cdot \ell_{i,t} \in [0,1)$.

4. By substituting the right term in the inequality of 3. for the e-term in 2. we get:
$$\frac{\sum_{i=1}^{n} c_{i,t+1}}{\sum_{i=1}^{n} c_{j,t}} \quad \leq \quad \sum_{i=1}^{n} w_{i,t} \cdot \left(1 - \eta \cdot \ell_{i,t} + \frac{\eta^2 \cdot \ell_{i,t}^2}{2}\right)$$
and by arithmetic transformation:
$$\leq \quad \sum_{i=1}^{n} w_{i,t} - \left(\eta \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}^2)\right)$$
By the normalisation of w: $\sum_{i=1}^{n} w_{i,t} = 1$, so:
$$\leq \quad 1 - \left(\eta \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}^2)\right)$$
By taking the ln on both sides of the inequality:
$$\ln\left(\frac{\sum_{i=1}^{n} c_{i,t+1}}{\sum_{i=1}^{n} c_{j,t}}\right) \quad \leq \quad \ln\left(1 - \left(\eta \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}^2)\right)\right)$$

5. By the inequality $e^{-x} \geq 1 - x$ (valid for any x) we get $\ln(e^{-x}) \geq \ln(1-x)$ and hence $-x \geq \ln(1-x)$. So, as an instance:
$$-\left(\eta \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}^2)\right) \quad \geq$$
$$\ln\left(1 - \left(\eta \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} (w_{i,t} \cdot \ell_{i,t}^2)\right)\right)$$
Verify that due to the assumptions in 1. $0 \leq \eta < 1$, the boundedness of loss ℓ by $[0,1]$ (axiom 2.2), as well as the normalisation of w our instance of x is within $[0,1]$.

6. By substituting the left (upper) term in the inequality of 5. for the right term in the inequality in 4. we get:

$$\ln\left(\frac{\sum_{i=1}^{n} c_{i,t+1}}{\sum_{i=1}^{n} c_{j,t}}\right) \leq -\left(\eta \cdot \sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t}) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t}^2)\right)$$

and by arithmetic transformation:

$$\leq \frac{\eta^2}{2} \cdot \underbrace{\sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t}^2)}_{\leq 1} - \eta \cdot \sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t})$$

… due to $\sum_{i=1}^{n} w_{i,t} = 1$, and $\ell \in [0,1]$, so:

$$\leq \frac{\eta^2}{2} \cdot 1 - \eta \cdot \sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t})$$

7. So, we arrived at the inequality (from 6.):

$$\ln\left(\sum_{i=1}^{n} c_{i,t+1}\right) - \ln\left(\sum_{i=1}^{n} c_{j,t}\right) \leq \frac{\eta^2}{2} - \eta \cdot \sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t})$$

Now we can sum up each side of the inequality from 1 to T:

$$\underbrace{\sum_{t=1}^{T}\left(\underbrace{\ln\left(\sum_{i=1}^{n} c_{i,t+1}\right)}_{=_{def} C_{t+1}} - \underbrace{\ln\left(\sum_{i=1}^{n} c_{j,t}\right)}_{=_{def} C_t}\right)}_{= (C_{T+1}-C_T)+\underset{=C_{T+1}-C1}{\cdots}+(C_3-C_2)+(C_2-C_1)} \leq \underbrace{\sum_{t=1}^{T}\left(\frac{\eta^2}{2} - \eta \cdot \sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t})\right)}_{=\frac{T \cdot \eta^2}{2} - \eta \cdot \sum_{t=1}^{T}\sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t})}$$

So, we arrive at:

$$\ln\left(\sum_{i=1}^{n} c_{i,T+1}\right) - \ln \underbrace{\left(\sum_{i=1}^{n} c_{i,1}\right)}_{=n \text{ by definition } 3.34} \leq \frac{T \cdot \eta^2}{2} - \eta \cdot \sum_{t=1}^{T}\sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t})$$

Hence:

$$\ln\left(\sum_{i=1}^{n} c_{i,T+1}\right) - \ln(n) \leq \frac{T \cdot \eta^2}{2} - \eta \cdot \sum_{t=1}^{T}\sum_{i=1}^{n}(w_{i,t} \cdot \ell_{i,t})$$

Recall, $c_{i,t}$ is the cumulative loss up to t in the exponent and we are after the bound for the regret with respect to the best predictor, hence we concentrate on the predictor with minimal cumulative loss up to T: Let us denote this predictor with 'b' ($b = (\iota i)(\sum_{t=1}^{T} \ell_{i,t} = min(\sum_{t=1}^{T} \ell_{1,t}, \ldots, \sum_{t=1}^{T} \ell_{n,t}))$). If there are more, then we can randomly pick one. Now:

$$\ln(c_{b,T}) \;\leq\; \ln\left(\sum_{i=1}^{n} c_{i,T+1}\right)$$

Hence:

$$\ln(c_{b,T}) - \ln(n) \;\leq\; \frac{T\cdot\eta^2}{2} - \eta\cdot\sum_{t=1}^{T}\sum_{i=1}^{n}(w_{i,t}\cdot\ell_{i,t})$$

8. By definition of c (definition 3.34):

$$c_{b,T} = \underbrace{c_{b,1}\cdot\prod_{t=2}^{T} e^{-\eta\cdot\ell_{b,t}}}_{\substack{=e^{-\eta\cdot(\ell_{b,1}+\ell_{b,2}+\cdots+\ell_{b,T})} \\ =\exp\left(-\eta\cdot\sum\limits_{t=1}^{T}\ell_{b,t}\right)}}$$

So:

$$\ln(c_{b,T}) = \ln\left(e^{-\eta\cdot\sum\limits_{t=1}^{T}\ell_{b,t}}\right) = -\eta\cdot\sum_{t=1}^{T}\ell_{b,t}$$

By substituting the right term in the last inequality in 7. we get:

$$-\eta\cdot\sum_{t=1}^{T}\ell_{b,t} - \ln(n) \;\leq\; \frac{T\cdot\eta^2}{2} - \eta\cdot\sum_{t=1}^{T}\sum_{i=1}^{n}(w_{i,t}\cdot\ell_{i,t})$$

And by arithmetical transformation:

$$\sum_{t=1}^{T}\sum_{i=1}^{n}(w_{i,t}\cdot\ell_{i,t}) - \sum_{t=1}^{T}\ell_{b,t} \;\leq\; \frac{T\cdot\eta}{2} + \frac{\ln(n)}{\eta}$$

If we substitute for η in accordance with 1: $\eta = \sqrt{\frac{2\cdot\ln(n)}{T}}$, we get:

$$\sum_{t=1}^{T}\sum_{i=1}^{n}(w_{i,t}\cdot\ell_{i,t}) - \sum_{t=1}^{T}\ell_{b,t} \;\leq\; \sqrt{2\cdot\ln(n)\cdot T}$$

Now, what is left is to employ the grey marked term for proving a bound for the meta-inductive method's regret.

9. According to definition 3.34, f_{emi} predicts as follows: $f_{emi,t} = \sum\limits_{i=1}^{n} w_{i,t} \cdot f_{i,t}$. Hence its loss is: $\ell\left(\sum\limits_{i=1}^{n}(w_{i,t} \cdot f_{i,t}), y_t\right)$. And hence its cumulative loss is:

$$\sum_{t=1}^{T} \ell\left(\sum_{i=1}^{n}(w_{i,t} \cdot f_{i,t}), y_t\right)$$

Since ℓ is convex (according to 1.), we get:

$$\ell\left(\sum_{i=1}^{n}(w_{i,t} \cdot f_{i,t}), y_t\right) \quad \leq \quad \sum_{i=1}^{n}(w_{i,t} \cdot \ell(f_{i,t}, y_t))$$

(I.e.: The loss of a weighted average of predictions is smaller than or equal to the weighted average of the losses of the predictions.) Note that '$\ell_{i,t}$' is just a short form for '$\ell(f_{i,t}, y_t)$'. Hence, from the last inequality in 8. and the convexity of ℓ we get:

$$\underbrace{\sum_{t=1}^{T}\left(\ell\left(\sum_{i=1}^{n}(w_{i,t} \cdot f_{i,t}), y_t\right)\right) - \sum_{t=1}^{T} \ell_{b,t}}_{aregret_{\langle emi,b\rangle,T}} \quad \leq \quad \sqrt{2 \cdot \ln(n) \cdot T}$$

Since f_b was the method with least cumulative loss up to T (we defined b this way in 7.), this regret bound holds also with respect to all other predictors.

□

Note that theorem 3.35 is analogous to thrm.6.9(i) in (Schurz 2019) which is based on theorems 2.2 and 2.3 of (Cesa-Bianchi and Lugosi 2006, pp.16f). These results prove a better bound than the one stated above, namely an absolute regret $\leq \sqrt{\ln(n) \cdot T/2}$.

Now, since according to theorem 3.35 $succ_{emi,t} \geq 1 - \frac{\sqrt{2 \cdot \ln(n) \cdot t}}{t}$, the term which is subtracted from 1 grows sublinearly ($1/\sqrt{t}$) with t and hence it seems as we have a no-regret algorithm. However, note also that the regret bound holds only for a learning parameter $\eta = \sqrt{\frac{2 \cdot \ln(n)}{T}}$ with a fixed T. So, we cannot simply calculate the limit of $succ_{emi,t}$, because η is not well defined for $T = \infty$. This is the reason why theorem 3.35 provides only a bound for regret up to arbitrary high T, a regret bound^T so to say. Since such a T represents a line separating the relevance of predictions, it is also called a *horizon* in online learning (see Cesa-Bianchi and Lugosi 2006, p.15). In long run optimisation we cannot draw such a line separating relevant from no longer relevant predictions, so we need to get rid of the horizon-dependency of the optimality result in theorem 3.35. This brings us to the second step of proving no-regret learnability with f_{emi}. It turns out that we can get rid of the horizon-dependency with only slight extra costs for the short run. The idea is that we divide time or rounds into periods that become increasingly longer, and that we start for each period the algorithm with the length of the period as the horizon. Since it is standard to do so by just doubling the length of the periods, this method is also called the *doubling trick* (see Mohri, Rostamizadeh, and Talwalkar 2012, p.158; see also Cesa-Bianchi and Lugosi 2006, p.17). We start with a horizon of 1 (round 1). Then we start the algorithm again with a horizon of 2 (rounds 2,3). Then we go on with a horizon of 4 (rounds 4,5,6,7). Then the horizon is 8 (rounds 8–15) and so on. So, the periods are $[2^m, 2^{m+1} - 1]$ and have length 2^m for $m \in \mathbb{N}$. The division of prediction rounds into such periods is depicted in figure 3.5. For each period we choose a learning parameter in which T is replaced by the horizon of the period, i.e. 2^m: $\eta_m = \sqrt{\frac{2 \cdot \ln(n)}{2^m}}$. The cumulative loss or the regret received until some time t is the sum of the cumulative losses or regrets of these periods.

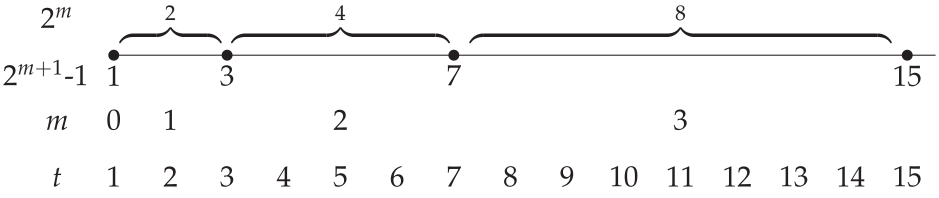

Figure 3.5: Doubling trick: An algorithm with bounds for some horizon is ran for increasing horizons by doubling the respective former horizon. The rounds are represented by t. The periods by m. At each $2^{m+1} - 1$th round ($m \in \mathbb{N}$) a new period starts. The length of such a period is 2^m rounds.

We can show that with such an "automatic update" of the horizon in the learning parameter, the following regret bound hold of f_{emi}:

Theorem 3.36 (Regret Bound for Exponential Weighting Algorithm)**.** *Given a loss function ℓ which is convex in its first argument, it holds for all prediction games G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ ($|\mathcal{F}| = n$): For all $1 \leq i \leq n$:*

$$\mathit{aregret}_{\langle emi,i \rangle,t} \quad \leq \quad \frac{2}{\sqrt{2} - 1} \cdot \sqrt{\ln(n) \cdot t}$$

Proof. In this proof we make use of the doubling trick as described above. The proof is a modification of a proof of (see Mohri, Rostamizadeh, and Talwalkar 2012, p.159) for our bound. Let us refer to (one of) the best predictor(s) in a period m by b_m. Then, by theorem 3.35, we get for each period with suitable learning parameter $\eta_m = \sqrt{\frac{2 \cdot \ln(n)}{2^m}}$:

$$aregret_{\langle emi, b_m \rangle, 2^{m+1}-1} \leq \sqrt{2 \cdot \ln(n) \cdot 2^m}$$

So, e.g., if we consider figure 3.5, it holds:

- $aregret_{\langle emi, b_0 \rangle, 1} \leq \sqrt{2 \cdot \ln(n) \cdot 2^0}$
- $aregret_{\langle emi, b_1 \rangle, 3} \leq \sqrt{2 \cdot \ln(n) \cdot 2^1}$
- $aregret_{\langle emi, b_2 \rangle, 7} \leq \sqrt{2 \cdot \ln(n) \cdot 2^2}$
- $aregret_{\langle emi, b_3 \rangle, 15} \leq \sqrt{2 \cdot \ln(n) \cdot 2^3}$

Now, at each $2^{m+1} - 1$th round a new period starts. So, for $t = 2^{m+1} - 1$, the regret of f_{emi} with respect to the best predictor(s) is at most the sum of the regrets of the periods before (it is equal if $b_m = b_{m-1} = \cdots = b_0$ and it is less otherwise):

$$\begin{aligned}
aregret_{\langle emi, b_m \rangle, t} &\leq \sum_{u=0}^{m} aregret_{\langle emi, b_u \rangle, 2^{u+1}-1} \\
&\leq \sum_{u=0}^{m} \sqrt{2 \cdot \ln(n) \cdot 2^u} \\
&\leq \sqrt{2 \cdot \ln(n)} \cdot \sum_{u=0}^{m} \underbrace{\sqrt{2^u}}_{2^{\frac{u}{2}} = (2^{\frac{1}{2}})^u = \sqrt{2}^u}
\end{aligned}$$

Note that $\sum_{u=0}^{m} \sqrt{2}^u$ is the geometric sum, so we get:

$$\begin{aligned}
&\leq \sqrt{2 \cdot \ln(n)} \cdot \sum_{u=0}^{m} \sqrt{2}^u = \sqrt{2 \cdot \ln(n)} \cdot \frac{\sqrt{2}^{m+1} - 1}{\sqrt{2} - 1} = \\
&\qquad = \frac{\sqrt{2 \cdot \ln(n)}}{\sqrt{2} - 1} \cdot \colorbox{lightgray}{$\left(\sqrt{2}^{m+1} - 1\right)$}
\end{aligned}$$

Note that $\sqrt{2}^{m+1} - 1 = \sqrt{2^{m+1}} - 1$.

Now, $\sqrt{2^{m+1}} - 1 \leq \sqrt{2 \cdot 2^{m+1} - 2} = \sqrt{2} \cdot \sqrt{2^{m+1} - 1}$

Recall that $t = 2^{m+1} - 1$, hence $\sqrt{2^{m+1}} - 1 \leq \colorbox{lightgray}{$\sqrt{2} \cdot \sqrt{t}$}$

By combining these inequalities (grey marked terms):

$$\leq \frac{\sqrt{2 \cdot \ln(n)} \cdot \sqrt{2} \cdot \sqrt{t}}{\sqrt{2} - 1} = \underbrace{\frac{\sqrt{2} \cdot \sqrt{2 \cdot \ln(n)}}{\sqrt{2} - 1}}_{\text{General form of doubling: } \frac{\sqrt{2}}{\sqrt{2}-1} \cdot bound(t)}$$

So, we get the bound:

$$aregret_{\langle emi,b_m\rangle,t} \quad \leq \quad \frac{2}{\sqrt{2}-1} \cdot \sqrt{\ln(n) \cdot t}$$

□

Note that the regret bound in theorem 3.36 is independent of any prediction horizon. If we compare the horizon-dependent bound in theorem 3.35 with the horizon-independent bound in theorem 3.36 we see that our (doubling) trick comes with the extra cost of a factor $\sqrt{2}/(\sqrt{2}-1) \approx 3.41$. However, this factor rapidly vanishes with increasing t.

Note also that Cesa-Bianchi and Lugosi (thrm.2.2 and 2.3 in 2006, pp.16f) and Schurz (thrm.6.9(ii) in 2019) prove a better bound for absolute regret, namely $\sqrt{2} \cdot \sqrt{\ln(n) \cdot t} + \sqrt{\ln(n)/8}$.

There is also a result in the online learning literature which shows a lower regret bound which cannot be optimised further, namely $\sqrt{\ln(n) \cdot t} \cdot 1/\sqrt{2}$ (see Cesa-Bianchi and Lugosi 2006, p.62; and the discussion in Schurz 2019, prop.6.14, sect.6.8). This means that in principle the short run regret can be optimised even further by about 85%. However, these algorithms are much more complicated and since we are after long run optimisation, we can stick to our relatively simple exponentially weighted learning algorithm.

Since the bound in theorem 3.36 is independent of any prediction horizon, we can now derive the main optimality result relevant for our epistemic endeavour:

Theorem 3.37 (Optimality of Exponential Weighting)**.** *In any regression game G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ (with $|\mathcal{F}| = n$, i.e. $\mathcal{F}$ is finite) it holds for all $1 \leq i \leq n$:*

$$\lim_{t\to\infty}(succ_{emi,t} - succ_{i,t}) \geq 0$$

Hence: f_{emi} is access optimal in the long run in G.

Proof. From theorem 3.36 we get for all $1 \leq i \leq n, t \in \mathbb{N}$:

$$\lim_{t\to\infty}\left(aregret_{\langle emi,i\rangle,t}\right) \leq 0$$

Note that by definition of $aregret$ and $succ$ it holds:

$$succ_{emi,t} = \frac{t - aregret_{\langle emi,i\rangle,t} - \sum\limits_{u=1}^{t} \ell_{i,u}}{t} = 1 - \frac{aregret_{\langle emi,i\rangle,t}}{t} - \frac{\sum\limits_{u=1}^{t} \ell_{i,u}}{t}$$

Also from the definition of *succ* we get:

$$succ_{i,t} = \frac{t - \sum_{u=1}^{t} \ell_{i,u}}{t} \quad \text{hence} \quad \sum_{u=1}^{t} \ell_{i,u} = t - t \cdot succ_{i,t}$$

By substitution we get:

$$succ_{emi,t} = 1 - \frac{aregret_{\langle emi,i \rangle,t}}{t} - \frac{t - t \cdot succ_{i,t}}{t} = succ_{i,t} - \frac{aregret_{\langle emi,i \rangle,t}}{t}$$

Hence:

$$succ_{emi,t} - succ_{i,t} = -\frac{aregret_{\langle emi,i \rangle,t}}{t}$$

And hence:

$$\lim_{t \to \infty} (succ_{emi,t} - succ_{i,t}) \geq 0$$

□

By this it follows immediately:

Theorem 3.38 (Possibility of No-Regret Regression)**.** *There is a no-regret learning algorithm f_l which is not para-scientific. Furthermore:*
Any finite *hypothesis set $\mathcal{F}$ of a regression game G is online learnable in the relative sense.*

In our applications later on we will sometimes use an algorithm which generates the weights directly out of the success rates instead of taking it (negative cumulative loss) in the exponent. Now, there is one technical detail relevant in defining such an algorithm: In order to achieve optimal performance, the learner needs to disregard those predictors who were outperformed by her already. In the case of exponentially weighting negative cumulative loss this is guaranteed since the influence of outperformed predictions vanishes exponentially ($\lim_{t \to \infty} \exp(-\sum_{u=1}^{t} \ell_{i,t}) = 0$ if the cumulative loss grows). Now, if we take the success rates directly and do not cut off outperformed predictions, their impact would not vanish and so the outperformed predictions prevent that the learner reaches the outperforming predictions. Figure 3.6 illustrates this problem and the cutting off solution by help of an example.

In order to avoid this problem, the learner needs to cut off (i.e. zero weight) those predictions which are already outperformed by it—note that this does not mean that a prediction method which was once outperformed by the learner gains never any influence again; this just means that such a method is ignored until it catches up with the learner again. We can implement *cutting off* by taking not the success rate $succ_{i,t}$ itself for calculating

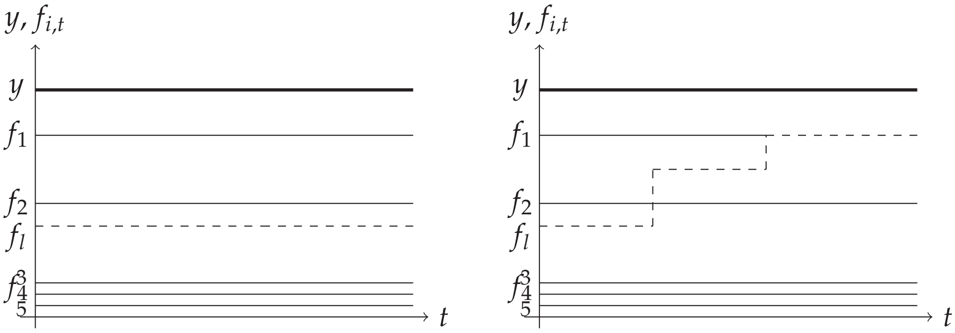

Figure 3.6: Example of taking success rates as weights without cutting off outperformed predictions (left) and with cutting off outperformed predictions (right): Ad left case: If the learner f_l just simply weights the predictions according to their success rates, the influence of predictions which are outperformed by f_l (here: f_3, f_4, f_5) might never vanish. This prevents that f_l reaches the better, outperforming prediction methods (here: f_1, f_2). Ad right case: If the learner cuts off the outperformed prediction methods, it reaches the better ones (first, by cutting off outperformed f_3, f_4, f_5, and then, by cutting off f_2 which is outperformed in the second round).

the weights, but $max(0, succ_{i,t} - succ_{l,t})$. Since the latter term expresses relative success of a prediction method f_i with respect to the learning method f_l or also how attractive f_i is for f_l, Schurz has also called this measure an *attractivity measure* (Schurz 2008b, p.296; and Schurz 2019, sect.6.6). Given this measure, we define a relative success or attractivity weighting meta-inductive learning method as follows:

Definition 3.39 (Attractivity Weighting)**.** We define relative success (attractivity) based weights recursively as follows (for all $1 \leq i \leq n, t \in \mathbb{N}$):

$$w_{i,1} = \frac{1}{n}$$

$$w_{i,t+1} = \begin{cases} \frac{max(0, succ_{i,t} - succ_{ami,t})}{\sum_{j=1}^{n} max(0, succ_{i,t} - succ_{ami,t})} & \text{if the denominator} > 0 \\ \frac{1}{n} & \text{otherwise} \end{cases}$$

Given these weights we define the relative success or attractivity based meta-inductive forecaster as:

$$f_{ami,t} = \sum_{i=1}^{n} w_{i,t} \cdot f_{i,t}$$

The regret bound for f_{ami} is as follows (regret grows also sublinearly) (see Schurz 2019, thrm.6.8):

Theorem 3.40 (Regret Bound for Attractivity Weighting Algorithm)**.** *Given a loss function ℓ which is convex in its first argument, it holds for all prediction games G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ ($|\mathcal{F}| = n$): For all $1 \leq i \leq n$:*

$$aregret_{\langle ami,i\rangle,t} \leq \sqrt{n \cdot t}$$

Proof. For a proof see (Cesa-Bianchi and Lugosi 2006, pp.12f, corollary 2.1 with the polynomial-parameter $p = 2$; and Schurz 2019, thrm.6.8). □

Note that according to the regret bound proven for the exponentially weighting meta-inductivist f_{emi} in theorem 3.36 $(2/(\sqrt{2}-1) \cdot \sqrt{\ln(n) \cdot t})$ and that of the attractivity weighting meta-inductivist f_{ami} in theorem 3.40 $(\sqrt{n \cdot t})$, f_{ami} is guaranteed to fare better than f_{emi} up to $n < 110$.

Recapping our results of the logic of deceivability for online regression we have:

- If a regression game G with $\mathcal{F}$ is realisable, then $\mathcal{F}$ is absolutely online learnable, if $\mathcal{F}$ (or its Littlestone's dimension) is finite.
- For any regression game (may it be realisable, a best expert or an agnostic game) G with $\mathcal{F}$ it holds: $\mathcal{F}$ is relatively online learnable, if $\mathcal{F}$ is finite.

This result about the impossibility of deceiving with respect to relative online learnability is very strong and allows for several epistemological applications. However, it only holds for the case of online regression, i.e. predictions with continuous values. For online classification, i.e. predictions with discrete values, the impossibility result regarding relative online learnability in an agnostic setting still stands. Since we cannot rule out such a setting from the outset, we are going to discuss several modifications which allow also for relative learnability in the case of online classification in the next chapter.

Chapter 4

Further Optimality Results

Here it is shown how the problem of suboptimality in online classification can be overcome by allowing for predictions in a modified setting. First, there is the possibility to randomise and by this achieve expected access optimality. Second, there is the possibility to predict as a group and achieve as a group arbitrary close access optimality. Finally, one can also put forward synchronisation constraints for online classification and online regression which rule out suboptimality in the case of classification. At the end the main optimality results of this part of the book are summarised.

In online classification Cover-style examples show that without any restriction, no prediction method is guaranteed to be access optimal: An adversary (🐱) can always design a prediction game such that the prediction method does not score at all, whereas the long run success rate of at least one learner is greater than zero. This is an impossibility result about relative learnability in the case of online classification.

Now, in the spirit of epistemic engineering as outlined in section 1.4, impossibility results are no *dead ends*, but points of departure: Since we know that in online classification an adversary is free to err us (recall that in contrast to this in online regression the adversary's hands are tied), we need to look out for new epistemic ends.

In order to overcome the problem of suboptimal online classification, three solutions are suggested in the literature. One is the randomisation approach according to which a classification meta-method should predict randomly, but with a bias towards the continuous meta-method's prediction (see Shalev-Shwartz and Ben-David 2014, sect.21.2). Another approach is the theory of collective weighted-average meta-induction which introduces a collective of meta-level methods whose average is aproximally access optimal (proposed in Schurz 2008b, sect.8). A further approach consists in enriching the formal structure of the problem by combining a discrete and a continuous prediction setting. The former is about qualitative belief; the latter about degrees of belief. To keep both systems synchronised one suggests synchronisation principles according to which qualitative and quantitative belief should be bridged; finally this bridging is intended to exclude adversarial scenarios in the case of online classification (see Feldbacher-Escamilla 2017b, sect.4).

In the following we are going to describe these approaches and discuss problems related to them. We start with the randomisation approach and show how the optimality result for online regression can be cashed out for proving expected success optimality (section 4.1). Afterwards, we describe in detail the approach to discrete predictions by help of collective action (section 4.2). We expand the discussion by an investigation of the problem of optimisation in a synchronised continuous and discrete prediction setting (section 4.3). Finally, we provide an overview of the results collected and achieved in part I of the book (section 4.4).

4.1 Classification and Randomisation

The randomisation approach is common in online learning and tries to overcome the gap between access optimality in a continuous and discrete setting via randomly picking out a prediction in such a way that the outcome is still biased towards an access optimal prediction method (see Cesa-Bianchi and Lugosi 2006, chpt.4).

In order to employ the optimality results of online regression as described in the preceding chapter also for the case of classification, we can reframe the case of online classification as follows (see Shalev-Shwartz and Ben-David 2014, pp.252f): We do not allow the adversary (😈) to set the true value after receiving the learner's prediction, because then Cover-style impossibility applies. Rather, we assume that the adversary has to set the true value before she receives the learner's prediction. Only in this way impossibility can be avoided. Now, we strengthen the case for the adversary by supposing that she has all information about the learner's prediction method at hand. So, the adversary knows the algorithm used by the learner. If in this case the learner's prediction strictly depends on the past and present predictions and outcomes, the adversary could calculate the learner's present prediction (i.e. make a true prediction about the learner's prediction) and by this err the learner again Cover-style. For this reason the learner's prediction is not allowed to strictly depend on past and present predictions and outcomes of $\mathcal{F}$ and $\mathcal{Y}$. In order to comply with this constraint, it is common practice to make the learning algorithm dependent on past and present predictions only in a non-strict, randomised way. This can be achieved as follows: Analogously to the case of online regression with the weighting methods f_{emi} and f_{ami}, the randomised learner calculates relative success dependent weights $w_{i,t}$ for each $f_i \in \mathcal{F}$ and round t. However, in contrast to online regression where the learning algorithm was a mixing of the f_is' predictions by taking the weighted average $\sum_i w_{i,t} \cdot f_{i,t}$, in online classification the learner cannot mix and so needs to decide for one of the predictions. Which prediction the learner chooses is decided randomly, but biased towards the weights: The learner is supposed to predict accordingly with $f_{i,t}$ with the probability of $w_{i,t}$. So, we can define a randomised online classification algorithm f_{rmi} as follows:

Definition 4.1 (Randomised Weighting)**.** We define relative success (attractivity) based weights recursively as in definition 3.39 as follows (for all $1 \leq i \leq n, t \in \mathbb{N}$):

$$\begin{aligned} w_{i,1} &= \frac{1}{n} \\ w_{i,t+1} &= \begin{cases} \frac{max(0, succ_{i,t} - succ_{rmi,t})}{\sum\limits_{j=1}^{n} max(0, succ_{i,t} - succ_{rmi,t})} & \text{if the denominator} > 0 \\ \frac{1}{n} & \text{otherwise} \end{cases} \end{aligned}$$

Given these weights we define the randomised attractivity based meta-inductive classificatory forecaster f_{rmi} as:

$$Pr(f_{rmi,t} = f_{i,t}) = w_{i,t}$$

Where $Pr(f_{rmi,t} = f_{i,t})$ is the probability of f_{rmi} to predict accordingly with f_i.

Note that for all rounds t there is an $f_i \in \mathcal{F}$ such that : $f_{rmi,t} = f_{i,t}$. So, the randomised meta-inductive forecaster f_{rmi} is not mixing. Note also that for any round t the adversary is allowed to know this probability of f_{rmi}. However, what the adversary is not allowed to know beforehand is f_{rmi}'s prediction, i.e. she has to set y_t before she receives $f_{rmi,t}$. Now, since the prediction of f_{rmi} at a round is randomised, we cannot say much about the bounds of her actual regret. However, since we have the probabilistic information about her predictions at hand, we can make a statement about the learner's expected regret. Since $Pr(f_{rmi,t} = f_{i,t})$ is within $[0,1]$, the bound of expected regret is just a specific instance of online learning regression (see thrm.6.10 and the proof in appendix 12.25 of Schurz 2019; the proof is based on Cesa-Bianchi and Lugosi 2006, sect.4.1f, but Schurz' proof is more explicit):

Theorem 4.2 (Regret Bound for Randomised Weighting Algorithm)**.** *Given a loss function ℓ which is convex in its first argument, it holds for all prediction games G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ ($|\mathcal{F}| = n$): For all $1 \leq i \leq n$:*

$$\mathbb{E}[aregret_{\langle fmi,i \rangle,t}] \;\leq\; \sqrt{n \cdot t}$$

Proof. For an exact proof see (Schurz 2019, theorem 6.10 plus appendix 12.25). □

Since $\mathbb{E}[succ_{fmi,t}] = \sum_{i=1}^{n} succ_{i,t} - \frac{\mathbb{E}[aregret_{\langle fmi,i\rangle,t}]}{t}$ we immediately get as a result the expected access optimality as characterised in definition 2.17:

Theorem 4.3 (Expected Optimality of Randomised Weighting)**.** *Given a loss function ℓ which is convex in its first argument, it holds for all prediction games G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ ($|\mathcal{F}| = n$): For all $1 \leq i \leq n$:*

$$\lim_{t\to\infty}(\mathbb{E}[succ_{fmi,t}] - succ_{i,t}) \geq 0$$

I.e.: f_{rmi} is expected to be access optimal in the long run.

In the following we are going to spell out how an randomisation might be implemented. The main idea is to interpret the predicted probabilities as frequencies. If, e.g., there are only two candidate methods, both of them gaining equal weights and one predicts 1 and the other 0, then the learning algorithm would predict 0.5 in an online regression game. However, in (binary) online classification it has to decide for 0 or 1. In the frequentist implementation we spell out here the learner might, e.g., predict 1 in the first case of this type and in the second 0 (and in further cases of this type the algorithm goes on to oscillate); by this the frequency of her prediction of cases of this type approaches the "ideal" prediction of 0.5 (see Feldbacher-Escamilla 2017b, sect.3.1).

It is important to keep in mind that the adversary is not allowed to know whether the learner uses this randomisation scheme, because otherwise she could again calculate the value $f_{rmi,t}$ before receiving it from f_{rmi} at t. Now, the idea of implementing randomisation is as follows: In predicting an event outcome one does not consider only past event outcomes, but all possibilities of past and present event outcomes; then one defines a prediction method that—regarding the binary setting—randomly predicts 0 or 1, but is—regarding a continuous setting—biased towards the ideal calculated value of the continuous setting. So, averaging over all possibilities, the method approaches the ideal calculated value in the finite case and reaches it in the long run. The details are as follows—this presentation is in accordance with (Schurz 2019, chpt.6.7.1): In order to explain the randomisation approach in detail, we expand the prediction setting further by the following elements:

- $\mathbb{Y}_1, \mathbb{Y}_2, \ldots$ is an infinite series of an infinite series of events; we identify $\mathbb{Y}_1$ with the infinite series of events above: $\mathbb{Y}_1 = Y_1, Y_2, \ldots$; and we use sub-sub-indices to pick out specific events: $\mathbb{Y}_{1_1} = Y_1$; analogously we refer to the outcome of the single events by $\mathbb{y}$, as, e.g., in

$\mathbb{y}_{1_1} = y_1$; finally, in the definitions of the score, success, and weight of an agents' prediction the series of events is always restricted to that provided in the argument place;

- $f_{rmi,t}$ is a qualitative prediction on event Y_t by a randomising meta-level agent.
  Note, since we need to compare and calculate the randomising learner's predictions for different event series, we will, if needed, make the respective event series in question explicit by writing $f_{rmi}(\mathbb{Y}_{u_t})$ instead of $f_{rmi,t}$ and relativising f_{rmi} to a different prediction game with event series $Y = \mathbb{Y}_u$.

The binary randomising meta-inductive agent f_{rmi} predicts within the limits of:

$$\mathbb{P}(f_{rmi}(Y_t) = 1) \approx f_{ami}(Y_t)$$

Here $\mathbb{P}(f_{rmi}(Y_t) = 1)$ is the ratio of the number of possible cases where f_{rmi} predicts 1 in round t and that of all $|\{0,1\}^t|$ possible series of binary event outcomes. Take, e.g., the outcomes of series of events as given in table 4.1 with the object-level predictions f_1 and f_2 (where $\mathbb{y}_1 = y$ is still considered to be the true series of outcomes, the other series of outcomes $\mathbb{y}_2$–$\mathbb{y}_8$ are the past outcomes, that are up to $t = 3$ possible; up to $t = 4$ there are 16 series possible, including the true outcome, etc.). Then a randomising meta-inductive method within the above stated limits would predict, e.g., according to table 4.2.

	$t = 1$	$t = 2$	$t = 3$	$\ldots$
$\mathbb{y}_6$	0	0	0	
$\mathbb{y}_2$	0	0	1	
$\mathbb{y}_3$	0	1	0	
$\mathbb{y}_4$	0	1	1	
$\mathbb{y}_5$	1	0	0	
$\mathbb{y}_1$	1	0	1	
$\mathbb{y}_7$	1	1	0	
$\mathbb{y}_8$	1	1	1	
f_1	1	1	1	1
f_2	0	0	0	0
f_{ami}	0.5	1.0	0.5	0.6

Table 4.1: Example of predictions of two object-methods (f_1, f_2) and one attractivity based weighting meta inductive method (f_{ami})

Of course there are numerous other randomising meta-inductive methods possible; important is only that their (weighted) average over all possible event series $\overline{f_{rmi}(\mathbb{Y})}$ approximates with increasing t the prediction made

	$t = 1$	$t = 2$	$t = 3$	...
$f_{rmi}(\mathbb{Y}_1)$	1	1	0	1
$f_{rmi}(\mathbb{Y}_2)$	1	1	0	1
$f_{rmi}(\mathbb{Y}_3)$	1	1	0	1
$f_{rmi}(\mathbb{Y}_4)$	1	1	0	1
$f_{rmi}(\mathbb{Y}_5)$	0	1	1	1
$f_{rmi}(\mathbb{Y}_6)$	0	1	1	0
$f_{rmi}(\mathbb{Y}_7)$	0	1	1	0
$f_{rmi}(\mathbb{Y}_8)$	0	1	1	0
$\overline{f_{rmi}(\mathbb{Y})}$	0.5	1.0	0.5	0.6

Table 4.2: Example of a randomised success-based meta-method: Such a method predicts in the binary case on average as often 1 as its real-valued prediction would be. So, e.g., given the real-valued predictions of table 4.1, it predicts in 50% of t_1-cases (where $f_1(Y_1) = 1$ and $f_2(Y_1) = 0$) 1 and in 50% of such cases 0. Analogously for all other cases. Which prediction the meta method makes in the end is chosen randomly/arbitrarily, but biased towards the real-valued prediction. For the optimality result important is the fact that the exact choice of the meta-method $f_{rmi}(\mathbb{Y}_{i_t})$ is probabilistically independent from the true outcome $\mathbb{Y}_{1_t}$.

according to attractivity weighted meta-induction (f_{ami}) better and coincides with it in the long run.

Now, assume that the pattern of the binary sequences in $1 \leq t \leq 3$ goes on this way; as can be seen in the tables above, a randomising meta-inductive method would not approach the best predictor's success rate in every possible event series. However, it is the randomised forecasters weaker aim of approaching the best predictor's success rate on average. Indeed, we can define a measure for expected success via the success of f_{rmi} with respect to an event series $\mathbb{Y}_k$ weighted by the probability of $\mathbb{Y}_k$ itself (which is a function of f_{rmi}'s predictions). The quite complicated formula for expected success in our implementation of randomisation is as follows (the big product produces the value for the probability of each event series $\mathbb{E}_k$):

$$\mathbb{E}[succ_{rmi,t}] =$$

$$\sum_{k=1}^{|\{0,1\}^t|} \prod_{l=1}^{t} (1 - f_{rmi}(\mathbb{Y}_{k_l}) - \mathbb{y}_{k_l} + 2 \cdot f_{rmi}(\mathbb{Y}_{k_l}) \cdot \mathbb{y}_{k_l}) \cdot succ_{rmi}(\mathbb{Y}_{k_t})$$

We have seen above, that $\mathbb{E}[succ_{rmi,t}]$ allows to for expected access optimality. Crucial for these bounds is an independence assumption stating that the true outcome and the prediction of the randomising meta-inductivist are probabilistically independent in the following way (see the independence assumption 6.10 in Schurz 2019):

$$\mathbb{P}(f_{rmi}(Y_t) = 1 | Y_t = 1 \;\&\; f_{ami}(Y_t) = r) \;=\; \\ \mathbb{P}(f_{rmi}(Y_t) = 1 | f_{ami}(Y_t) = r) \;\text{ for all } r \in [0,1]$$

This means that an adversary is not allowed to pick the y_ts for a series of events in such a way as if she knew already beforehand the predicted value $f_{rmi,t}$.

A nice feature of randomisation in the discrete settings is its structural closeness to the continuous case. However, considering the independence assumption above it is clear that a daemonic scenario is ruled out only by stipulation. Furthermore, the relativisation of the optimality result to expected predictive success ($\mathbb{E}[succ_{rmi,t}]$) instead of predictive success *per se* ($succ_{rmi,t}$) opens another dimension into the infinite whose trend is even opposed: Whereas in the continuous case strict access optimality is restricted to the long run, i.e. to infinite series of predictions, in the randomising approach access optimality is restricted to the long run as well as to weighted averaging among the set of possible outcomes; since the number of possible event outcomes increases with the number of predictions, in the long run, information about expected success decreases.

These drawbacks of optimality by help of randomised classification lead us to consider another proposal that is about access optimality of predictive success *per se* in a discrete setting.

4.2 Classification and Collective Action

In (Schurz 2008b, sect.8) a set of qualitative meta-inductive prediction methods is defined which, on average, transforms access optimality results for a continuous setting to the discrete realm. The idea is as follows: If one wants to approach a value of a continuum by help of discrete values, one may arrange discrete values around the value of the continuum in such a way that the average of the discrete values is close to the value of the continuum. E.g., one can approach/reach $0.5 \in [0,1]$ by averaging over the elements of $\{0,1\}$. Similarly for $0.75 \in [0,1]$ by averaging over elements of $\{0,1\}$: $0.75 = (0+1+1+1)/4$. Now, in a discrete setting, like the binary setting, only discrete predictions, e.g., binary predictions, are admissible. So, every method can predict only a value out of $\{0,1\}$. However, the number of prediction methods is in principle not fixed. This can be exploited by a meta-strategy by settling around the value of a continuum 0/1-predicting methods in such a way that on average the value of the continuum is approached. So, e.g., if the calculated ideal prediction is 0.25, then the meta-method can approach it by averaging over one 1-predictor and three 0-predictors: $0.25 = (1+0+0+0)/4$. In the binary prediction setting no meta-method can exploit this fact directly, because averaging over the predictions leaves the binary value space. However, on a meta-meta level where one can compare successes of object- and meta-methods as, e.g., we are doing, a meta-meta-method can average over the single prediction method's success and can exploit this on the meta-meta-level.

In order to indicate such a meta-meta-method, we add to the discrete prediction setting a group of binary meta-inductive methods:

- $f_{cmi_1,t}, \ldots, f_{cmi_k,t}$ are the qualitative predictions on Y_t of k meta-level agents

Now, Schurz (2008b) has found an interesting way of emulating real-valued success-based predictions in the discrete setting by defining the meta-inductive predictions as follows ($[\cdot]$ rounds to the next integer, as, e.g. $[0.75] = 1, [0.25] = 0, [0.5] = 1$):

Definition 4.4 (Collective Weighting)**.**

$$f_{cmi_i,t} = \begin{cases} 1 & \text{if } i \leq [f_{ami,t} \cdot k] \\ 0 & \text{otherwise} \end{cases}$$

So, if, e.g., $k = 10$ and the ideal (continuous) predicted value $f_{ami,t} = 0.75$, then the first seven meta-inductivists predict 1 ($1, \ldots, 7 \leq 0.75 \cdot 10$), and the remaining three meta-inductivists predict 0 ($8, \ldots, 10 > 0.75 \cdot 10$). By this a meta-meta-inductivist can exploit the meta-inductivists' predictions by averaging and approximating 0.75 by 0.7. In this case, using only a subset of four meta-inductivists would perform better. It turns out that, although each meta-inductivist's success rate is not bounded by the object-level methods' success rates, the average of them is (see Schurz 2008b, p.299):

Theorem 4.5 (Bound for Collective Weighting Algorithm)**.** *Given a loss function ℓ which is convex in its first argument, it holds for all prediction games G with the true values $\mathcal{Y}$ and the predictions $\mathcal{F}$ ($|\mathcal{F}| = n$): For all $1 \leq i \leq n$:*

$$\overline{succ}_{\{cmi_1,\ldots,cmi_m\},t} - succ_{i,t} \geq \sqrt{\frac{n}{t}} - \frac{1}{2 \cdot k}$$

Proof. For a proof see (Schurz 2008b, p.299; and Schurz 2019, sect.6.7.2, thrm.6.11 plus appendix 12.26). □

For the long run, i.e. the limiting case, the distance of the average shrinks as a function of the number of meta-level methods k to $1/(2 \cdot k)$. The trick of this collective average-weighting method is to mimic access optimality in the real-valued case, which is achieved by, lets say, $f_{ami,t}$, via bringing $\overline{f_{cmi,t}}$ as close as possible to $f_{ami,t}$. It is, so to say, lifting the binary predication game onto a meta-meta level of a prediction game with $1/(2 \cdot k)$ (with arbitrary high k) as the smallest approximatable unit.

Averaging success rates means that the meta-inductive agents $f_{cmi_1}, \ldots, f_{cmi_k}$ have to share their success. According to the general impossibility result regarding a daemonic setting one cannot define a meta-method which takes as input the predictions of $f_{cmi_1}, \ldots, f_{cmi_k}$ and produces a prediction on its own. So, one might say that in order to deal with the problem of discrete predictions from a meta-inductive perspective one is forced to act as a collective.

The main advantage of this approach is to be found in its applicability to any prediction setting whatsoever. One also does not have to exclude a daemonic scenario by stipulation, as is done in the randomisation approach. Although in such a setting all meta-inductivist methods might perform suboptimally, on average these methods approximate access optimal performance. However, it guarantees *approximation* of access optimality only in the long run. As we have seen, the lower bound of the average success rate is in the long run $max(succ_{1,t}, \ldots, succ_{n,t}) - 1/(2 \cdot k)$. Now, although k might be chosen arbitrarily high, one cannot achieve equal success rates in the long run. So, for some daemonic scenarios even the collective of meta-inductive methods will predict suboptimally.

To sum up the results of this and the preceding section, discrete prediction settings allow for daemonic scenarios; randomisation allows for weak access optimality in the sense of convergence of expected success rates with that one of the best object-level method accessible, but at cost of stipulating independence between meta-inductive prediction and true outcome, thus stipulating that daemonic scenarios are impossible. On the other hand, collective meta-induction allows for an approximation of average success as accurate as one wishes; however, strict convergence is not always possible and by this also a collective of meta-inductive methods performs suboptimal in at least some daemonic settings, even in the long run. This facts seem to suggest that in order to approach the problem of induction within a discrete setting one has to enrich the structure of the problem and try to prove meta-inductive access optimality or the impossibility of a daemonic setting for such an enriched structure. This is the line of argumentation we are following in the next section by considering the problem of daemonic settings within a synchronised prediction environment.

4.3 Classification and Synchronisation Constraints

Instead of focusing on a modification of the setting (randomisation or considering collective success), we are here concerned with linking the settings, namely the classification and the regression setting. The idea is that if we do not know whether nature allows also for continuous predictions or for discrete predictions only, we might want to provide both kinds of predictions. However, in providing two kinds of predictions of one and the same event, it seems plausible to assume that these two different kinds of predictions should be linked somehow. In this section we frame the problem as linking qualitative beliefs with quantitative ones. The idea is that we can put forward constraints for keeping both systems synchronised. Afterwards, we aim at showing that these constraints exclude cases in which relative online learning is suboptimal. The argumentation in this section is along the line of (see Feldbacher-Escamilla 2017b, sect.4).

We want to exclude Cover-style prediction games. Particularly we are after excluding daemonic scenarios with the following properties:

1. The object-method's success rates are limited: For all $1 \leq i \leq n$ there exists $\lim\limits_{t\to\infty} succ_{i,t}$.
2. The meta-inductive learning method performs long run suboptimal: There is an $1 \leq i \leq n$ such that: $\lim\limits_{t\to\infty}(succ_{mi,t} - succ_{i,t}) < 0$
   Note that by this characterisation it is supposed that also the success rate of the meta-inductivist or at least its upper bound with respect to the best performing method(s) is limited.

   (From these two assumptions it follows that at least one object-method is predictively successful.)

As we mentioned in the preceding section, we suggest structural enrichment for solving the problem of suboptimal success based predictions in a discrete setting. The structure we are interested in is a synchronised setting. So, we combine a discrete prediction setting with a continuous one and put forward some synchronisation constraints. We aim at showing that, given these constraints, daemonic scenarios are impossible. Daemonic scenarios underlay the meta-inductivists suboptimality. So, arguing for the impossibility of such daemonic scenarios in a synchronised setting is the same as to argue for the optimality of meta-induction in all reasonable synchronised settings.

Since we want to motivate the synchronisation constraints epistemically, we choose a formalism that allows for such an interpretation:

- The classification or discrete prediction setting consists of:
  - $Y_1, Y_2, \ldots$ an infinite series of binary events whose outcomes $y_1, y_2, \ldots$ are within $\{0, 1\}$
  - $Bel_1(Y_t), \ldots, Bel_n(Y_t)$ are the binary predictions concerning Y_t (elements of $\{0, 1\}$) of the n object-level agents

  - $Bel_{smi}(Y_t)$ which is the binary prediction concerning Y_t by the meta-inductive agent (who aims at "synchronisation", hence 's'—details see below)

- The regression or continuous prediction setting consists of:
  - The same series of binary events
  - $Pr_1(Y_t), \dots, Pr_n(Y_t)$ which are real-valued predictions concerning Y_t (elements of $[0,1]$) of the n *object-level agents*
  - $Pr_{smi}(Y_t)$ which is a real-valued prediction concerning Y_t of the *meta-level agent*

Now, *Bel* is interpreted as qualitative belief or acceptance in the sense that '$Bel_i(Y_t) = 1$' is supposed to mean that according to method i $y_t = 1$ or just simply: agent i believes that Y_t will take place; analogously '$Bel_i(Y_t) = 0$' means that agent i believes that Y_t will not take place. Note that we assume here that beliefs are complete in the sense that for every event Y_t the agent either believes that it will take place or believes that it will not take place. In principle one might try to relax the completeness condition by allowing agents to abstain from judgement. But then, of course, the question arises of how to adequately take into account abstention in scoring. We will stick to the idealisation of completeness, this the more since under specific circumstances expert knowledge also spreads from object-level methods to meta-level methods in a setting with restricted access only, where restricted access might be equalised with incompleteness (for details see Thorn and Schurz 2012, sect.7f).

Similarly we re-interpret *Pr*: '$Pr_i(Y_t) = r$' is now supposed to mean that i's degree of belief that Y_t will take place is r. Regarding scoring, this re-interpretation seems to be fine inasmuch as scoring can be directly related to betting behaviour. If outcomes are binary, it holds that the more an agent tends to extremes $(0,1)$, the higher are also her chances in scoring well. But at the same time also her risk is of not scoring at all—so scoring tends also to the extremes then. And the more an agent tends to the indefinite (0.5), the safer she scores, but also the smaller the scores. In the case of a constant degree of belief of 0.5, expected predictive success will be equal to a randomisation among all possible event series as, e.g., is the case of flipping a fair coin—which is, to say the least, not a remarkably good benchmark.

Now, we define the meta-inductive methods intended for synchronised predictions as follows:

Definition 4.6 (Synchronised Weighting)**.**

$$Pr_{smi}(Y_t) = f_{ami,t}$$

$$Bel_{smi}(Y_t) \begin{cases} 1 & \text{if } Pr_{smi}(Y_t) > 0.5 \\ 0 & \text{if } Pr_{smi}(Y_t) \leq 0.5 \end{cases}$$

Due to our optimality and suboptimality results, Pr_{smi} is long run access optimal, whereas Bel_{smi} is not. Given such an expanded structure, what rationality constraints can be put forward? In the light of the re-interpretation provided above it seems to be appropriate to put forward synchronisation principles between the qualitative and the quantitative series of beliefs. The first principle we think of is a synchronisation principle that acts event-wise between these systems. It is a very specific case of the so-called *Lockean thesis* and states that a degree of belief above a specific threshold is necessary and sufficient for qualitative belief or acceptance. Since we are dealing with complete qualitative belief, the *natural* threshold is 0.5. Otherwise the situation could arise that one qualitatively believes a proposition and disbelieves its negation, although one's degree of belief in the proposition is strictly lower than the degree of belief in the negation, which sounds at least paradoxical (see, e.g., Leitgeb's critique of Lin and Kelly's approach in Leitgeb 2017). Since this synchronisation principle is event-wise, we call it a 'synchronous synchronisation principle'. It is as follows:

$$Bel_i(Y_t) = \begin{cases} 1 & \text{if } Pr_i(Y_t) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{SynSync}$$

The case of $Pr_i(Y_t) = 0.5$ is, epistemically speaking, not clearly regulated—regarding complete belief one might belief or disbelief that Y_t will take place. For our argument below one can uphold a principle (SynSync*) similar to (SynSync) above, where $Pr_i(Y_t) = 0.5$ enforces one to set $Bel_i(Y_t) = 1$. What matters only is that all cases of $Pr_i(Y_t) = 0.5$ are treated the same way, i.e. enforce either $Bel_i(Y_t) = 0$ or $Bel_i(Y_t) = 1$. (SynSync) can be also expanded to discrete settings with more than two admissible qualitative predictions. If, e.g., there are three qualitative predictions admissible ($\{0, 0.5, 1\}$), then one could state that $Bel_i(Y_t) = 1$, if $Pr_i(Y_t) > 2/3$, $Bel_i(Y_t) = 0.5$, if $1/3 < Pr_i(Y_t) \leq 2/3$, and $Bel_i(Y_t) = 0$, if $Pr_i(Y_t) \leq 1/3$. However, such a general bridging between the quantitative and the qualitative realm needs further argumentation and so we are not going deeper in this matter; in the literature often so-called non-epistemic values are cited for such a bridging (see, e.g., Longino 2008). Note that the binary meta-inductive prediction method Bel_{smi} from above satisfies this constraint by definition. For all object-level methods in a daemonic scenario (SynSync) poses no problem, since they can easily pick out a (partial) probability function that satisfies for each event this constraint. In the daemonic example Cover-style mentioned above the agents with $Bel_1(Y_t) = 1$ and $Bel_2(Y_t) = 0$ might simply equate their constant qualitative belief with their quantitative one. A discussion of such a synchronisation constraint in the framework of prediction games is also provided in (Schurz 2019, sect.8.1.7, particularly the optimality principle 8.3).

Beside this constraint we suggest another one for diachronic considerations. The idea is as follows: An agent i might believe or disbelieve and have an event-wise synchronised degree of belief above or below the threshold 0.5 regarding the event's taking place or not. However, the event-wise synchronisation does not oblige i to synchronise her degrees of belief according to her qualitative predictions in the long run. Take, e.g., an agent i with the alternating acceptance behaviour and equal degrees of belief according to table 4.3. Although both epistemic attitudes towards Y_t are event-wise synchronous, one might ask whether it is rational for i to stick to her degrees of belief also in the long run or whether she should at some point in time s adopt her degrees of belief also according to her past prediction behaviour? We think that in order to be diachronic synchronous too, agents should also calibrate—although it should be highlighted that this is already a much stronger assumption than that of synchronous synchronisation.

	t=1	t=2	t=3	t=4	t=5	t=6	$\cdots$
$Bel_i(Y_t)$	1	0	1	0	1	0	$\cdots$
$Pr_i(Y_t)$	1.0	0.0	1.0	0.0	1.0	0.0	$\cdots$

Table 4.3: Example of qualitative and quantitative predictions

How should calibration in such a setting work? Of course we do not suggest to oblige the agent to calibrate directly according to the past outcomes—this would lead to a constraint of applying the straight rule. The straight rule is to be considered as a possibility for an object-level method; but it (or convergence in the limit with it) is not to be considered as a necessary condition for rationality (see our discussion in part II). On the other hand, calibration according to past predictions alone seems to be a constraint too weak to be upheld. It would not ground the agent's degrees of belief to (experimental) outcomes at all. In the example given above the agent was obliged to calibrate her degrees of belief in the long run, i.e. starting at a point in time s, to 0.5; and this regardless of the past outcomes. What seems to be more reasonable is to demand calibration with respect to one's own predictions *and* the past outcomes. Recall that according to the definition above, *success* combines both via keeping track of an agent's *true* predictions in comparison to *all predictions* made so far by her. For this reason we suggest as a middle ground between purely outcome-oriented calibration and calibration that is based on past predictions only: success-oriented calibration for diachronic synchronicity. So, we think that in the long run an agent's degrees of belief should be calibrated by her success rate in the following way:

There is an s, such that for all $r \geq s$:

$$\begin{aligned} Bel_i(Y_r) = 1 \quad &\Rightarrow \quad Pr_i(Y_r) \; = \; \lim_{t\to\infty} succ_{Bel_i,t} \\ Bel_i(Y_r) = 0 \quad &\Rightarrow \quad Pr_i(Y_r) \; = \; 1 - \lim_{t\to\infty} succ_{Bel_i,t} \end{aligned} \tag{DiaSync}$$

The principle (DiaSync) states about quantitative belief, in order to be diachronically synchronised, the following: There is a point in the series s such that for all events following Y_s, i.e. for all Y_r with $r \geq s$, the quantitative belief regarding Y_r's taking place ($Pr_i(Y_r)$) equals the limiting success rate regarding Bel_i. It is clear that (DiaSync) holds only, if the success rate is limited. This means that there is a point in the series where the success rate is fixed, where the object method reached an "equilibrium" regarding closeness of the predictions to the truth. The idea is that s is after such a limiting point.

(DiaSync) can be expanded also to a discrete setting where the number of admissible predictions is greater than two, not only in $\{0, 1\}$. So if, e.g., the admissible quantitative predictions are in $\{0, 0.5, 1\}$, then for $Bel_i(Y_r) = 1$ and $Bel_i(Y_r) = 0$ things may remain as in (DiaSync); and with respect to $Bel_i(Y_r) = 0.5$ the degree of belief in Y_r's taking place may be equalised with the value in-between them; the third value of such a discrete setting may then be plausibly interpreted as *suspension of judgement* (the outcome may be interpreted as undetermined). That there is always a plausible interpretation for a qualitative value in such an extended diachronic synchronisation principle is, of course, not guaranteed. But if there is a "bridge" between the qualitative and quantitative system under investigation, then it seems that one can also make sense of an extended diachronic synchronisation principle.

In our description of daemonic scenarios we have stipulated that in such scenarios the success rates of the object-methods are limited. So, (DiaSync) is supposed to hold for quantitative beliefs in such scenarios. If an agent believes that an event Y_r will take place, then her degree of belief in Y_r's taking place should cohere with her past performance in predicting Y-events. And if an agent believes that an event Y_r will not take place, then her degree of belief in Y_r's not taking place should—completeness of belief presupposed—equal the inverse of her degree of belief in Y_r's taking place. We have argued above that just considering the event outcomes in calibration would be inadequate since it would enforce the straight rule. Such a calibration principle might be considered as a *purely* empirical constraint. On the other hand, just calibrating according to one's past predictions seems to be without any empiricistic spirit at all. An agent would be considered diachronically rational if she just sticks to her method. In case the used method is *a priori*, also the calibration principle would lead from *a priori* predictions to *a priori* ones. Hence, one might consider such a principle as rationalistic in spirit. By stipulating diachronic coherence of predictions through equalising degrees of belief with limiting success rates (if they exist), we think one gets the right spin from both camps: One remains in an empirically informed way with one's method.

It should be mentioned here that the diachronic synchronisation principle is used not as a reflection principle in our argument. It is not intended that *de facto* an agent should be supposed to update her degrees of belief according to her success rate—since this information is available only for the limiting case such an application would be too much to ask for. However, we, talking about daemonic scenarios and having knowledge about the limiting case, may reasonably put forward constraints also for this case. And we think that from this perspective (DiaSync) is reasonable to ask for. Note that also the meta-inductive solution to the problem of induction holds strictly speaking only for the limiting case—only for this case it can be shown that the meta-inductive weighting method is among the best accessible methods within the setting (although, of course, the short run results demonstrate some kind of "epistemic controllability" by help of meta-induction). In order to uphold access-optimality (DiaSync) just adds another consideration to the limiting case: Meta-induction remains access-optimal also in a setting where discrete predictions are coupled with continuous ones, if all the agents within the setting are diachronically coherent, i.e. calibrated.

According to this proposal, the alternating predictions above would force an agent to calibrate her degrees of belief depending on the outcomes of the events as given in table 4.4. In the first case, predictions are in complete agreement with outcomes, so it seems to be plausible that an agent trusts completely in her prediction method (regarding qualitative belief) in the long run; analogously in the second case, where predictions are in complete disagreement with outcomes; here it seems to be plausible that an agent distrusts her prediction method (regarding qualitative belief) completely in the long run. Finally, in the third and fourth case, where just 50% of the predictions are correct, an agent should trust in her method (regarding qualitative belief) no more, but also no less, than trusting in flipping a fair coin. Note that the last two cases represent the object methods in our Cover-style example of a daemonic scenario (we just switched the values predicted by the methods with that of the event outcomes here).

Let us apply the framework presented above to the problem of daemonic scenarios. Now, as was pointed out above, in a daemonic setting the success rates of the relevant, i.e. the best, object-level agents converge. So, it holds:

$$\lim_{t\to\infty} succ_{Bel_1,t} \quad = \quad \lim_{t\to\infty} succ_{Bel_2,t}$$

But then, in order to be diachronically synchronised, the degrees of belief

	t=1	t=2	t=3	t=4	t=5	t=6	$\cdots$	$\cdots$
$Bel_i(Y_t)$	1	0	1	0	1	0	$\cdots$	$\cdots$
Y_t	1	0	1	0	1	0	$\cdots$	$\cdots$
$succ_{Pr_i,t}$	$\frac{1}{1}$	$\frac{2}{2}$	$\frac{3}{3}$	$\frac{4}{4}$	$\frac{5}{5}$	$\frac{6}{6}$	$\cdots$	1
Y_t	0	1	0	1	0	1	$\cdots$	$\cdots$
$succ_{Pr_i,t}$	$\frac{0}{1}$	$\frac{0}{2}$	$\frac{0}{3}$	$\frac{0}{4}$	$\frac{0}{5}$	$\frac{0}{6}$	$\cdots$	0
Y_t	1	1	1	1	1	1	$\cdots$	$\cdots$
$succ_{Pr_i,t}$	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{2}{4}$	$\frac{3}{5}$	$\frac{3}{6}$	$\cdots$	$\frac{1}{2}$
Y_t	0	0	0	0	0	0	$\cdots$	$\cdots$
$succ_{Pr_i,t}$	$\frac{0}{1}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{2}{5}$	$\frac{3}{6}$	$\cdots$	$\frac{1}{2}$

Table 4.4: An example of predictions and their corresponding success-rates showing that object-predictors in a daemonic scenario are not calibrated according to (DiaSync)

of the agents are also calibrated equally in the long run: By (DiaSync) we get for some point s in Y (in case s differs agent-wise one has to choose the "larger" one):

There is an s, such that for all $r \geq s$:

$$\begin{array}{rcrcl} Bel_1(Y_r) = 1 & \Rightarrow & Pr_1(Y_r) & = & \lim\limits_{t\to\infty} succ_{Bel_1,t} \\ & & & & \shortparallel \\ Bel_2(Y_r) = 0 & \Rightarrow & Pr_2(Y_r) & = & 1 - \lim\limits_{t\to\infty} succ_{Bel_2,t} \end{array}$$

Since in the binary case with two admissible predictions of the daemonic scenario the object-level agents' success rates are 0.5, by (SynSync) we get indiscernibility of qualitative beliefs, i.e. we get for some point s in Y that for all $r \geq s$:

$$Bel_1(Y_r) \quad = \quad Bel_2(Y_r)$$

But then the meta-level agent Pr_{smi} and her qualitative counterpart Bel_{smi} would at some point in Y predict exactly the same way as both object-level agents Pr_1, Bel_1 and Pr_2, Bel_2 do. So the object-level and the meta-level methods' success rates would converge which means that the setting cannot be a daemonic one.

This result also holds for a binary daemonic scenario with more than two object methods, since their success rates still have to converge in order to be attractive for the meta-inductive method; by this, again, their degrees of belief converge (DiaSync); and by this, again, their qualitative beliefs converge (SynSync). In case the admissible predictions are not binary, but discrete to a degree greater than 2, also daemonic scenarios are impossible. Consider, e.g., the case where $Bel_1(Y_t)$ is constantly 1, $Bel_2(Y_t)$ is constantly 0, and $Bel_3(Y_t)$ is constantly 0.5. Their success rates also have to converge and can be maximally 1/3. But then, by an expanded version of (DiaSync), $Pr_1(Y_r)$ (for all $r \geq$ some s) equals also a value $\leq 1/3$. However, this would be incoherent with an expanded version of (SynSync), enforcing, e.g.: $Pr_1(Y_r) > 2/3$.

To sum up our main argument against the possibility of a daemonic scenario in such synchronised settings runs as follows:

1. A daemonic setting with successful agents enforces different qualitative beliefs, but equal success rates in the long run among the relevant object-level agents. (our definition of a daemonic setting)

2. Equal success rates in the long run enforce equal calibration of degrees of belief. (see (DiaSync))

3. Equal calibration of degrees of belief enforces equal qualitative beliefs. (see (SynSync))

4. Hence, no daemonic setting satisfies synchronic and diachronic synchronisation constraints at the same time. (1–3)

So, in case of a richer structure of prediction tasks meta-inductive suboptimality can be overcome by putting forward synchronisation constraints: If all agents in the combined qualitative and quantitative setting are rational in the sense that they are synchronically and diachronically synchronised (calibrated), then no daemonic scenarios are possible and by this meta-induction remains access optimal.

As the three approaches to learnability and online classification in this chapter show, the impossibility result about relative learnability in online classification asks for setting different ends and engineering different means in order to achieve these ends. In the next section we summarise them together with the other main results about online learnability.

4.4 Summary of Main Results

Let us take stock of the results we have achieved so far: We can categorise online prediction games as follows: There are online classification games (discrete values) and there are online regression games (continuous values within $[0, 1]$). Both of them can be subdivided into realisable, best expert, and agnostic prediction games. A prediction game is realisable, if it contains in the predictor or hypothesis set $\mathcal{F}$ also the true series $\mathcal{Y}$. It is a best expert game, if there is one hypothesis which is never outperformed by any other prediction method (since the true series is never outperformed by any other prediction method, it follows that every realisable game is also a best expert game). If it is no best expert game, then it is agnostic. Figure 4.1 gives an overview about this different prediction games.

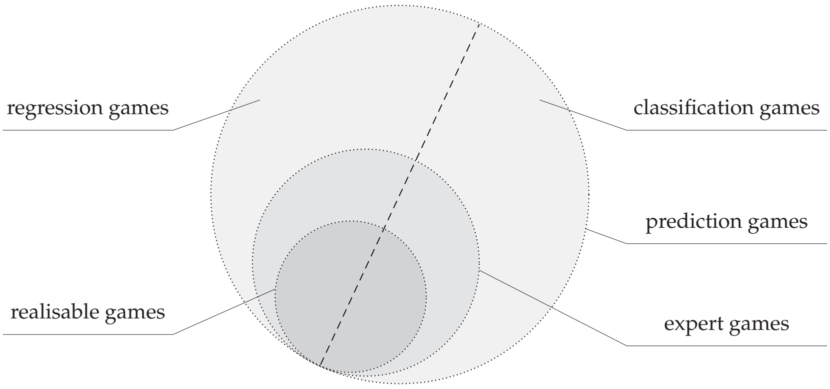

Figure 4.1: A taxonomy of prediction games

Now, let us come to some preconditions for the online learnability results we have gained in this part of the book:

- Regarding the loss ℓ:
  - Online classification: ℓ is the 0-1 loss (in case of multiclass classification with $k > 2$ one can also use a 0-k loss which penalises wrong predictions not fully, but by some value which increases with the distance in some ordering of the k values)
  - Online regression: ℓ is convex
- Regarding the predictor or hypothesis set $\mathcal{F}$:
  - Online classification: $Ldim(\mathcal{F})$ is finite
  - Online regression: $\mathcal{F}$ is finite

Now, given these conditions, we can summarise the possibility and impossibility results regarding online learnability as follows: Absolute online learnability is guaranteed only for realisable prediction games (possibility result). Best expert games or agnostic games allow not for guaranteed absolute online learnability (impossibility result). Regarding relative online learnability it matters whether the setting is one of online classification or one of online regression. Relative learnability in online classification is guaranteed only for best expert games (possibility result), but not for agnostic ones (impossibility result). One can bypass the impossibility result by allowing randomisation, collective action or synchronisation. Randomisation amounts to providing a probabilistic prediction and proves to be expected optimal. Collective action allows to approximate collective optimality, but, strictly speaking, does not achieve it. Finally, putting forward synchronisation allows for optimality, if all classificatory predictions are supposed to be synchronically and diachronically synchronised with their respective predictions of online regression. Relative learnability in online regression is also possible in agnostic prediction games (main possibility result). Table 4.5 provides an overview of these possibilities and impossibilities.

	absolute learnability $\lim\limits_{t\to\infty} succ_{l,t} = 1$	**relative learnability** $\lim\limits_{t\to\infty}(succ_{l,t} - succ_{i,t}) \geq 0$
online classification	realisable games	best expert games; agnostic games also, if one is allowed to randomise, act collectively (approximates optimality only) or in case of a synchronised setting.
online regression	realisable games	agnostic games

Table 4.5: Overview of possibility/impossibility results on absolute and relative online learnability: f_l is a learning algorithm, f_i is any method of $\mathcal{F}$.

The most important algorithms we will employ in the following are:

- f_{ami}: Attractivity based meta-inductive forecaster with a regret bound of $\sqrt{n \cdot t}$

- f_{emi}: Exponential meta-inductive forecaster with a regret bound of $2 \cdot \sqrt{\ln(n) \cdot t}/(\sqrt{2} - 1)$

Finally, let us state the assumptions for this optimality results again:

- The loss ℓ is bounded and convex (in online regression) or 0-1 (in online classification).

- The past and present predictions of $\mathcal{F}$ are complete and accessible to the learner, i.e. for each round t and $f_i \in \mathcal{F}$: $f_{i,\leq t}$ is accessible to the learning algorithm f_l.

- The past true outcomes (of $\mathcal{Y}$) are accessible to the learner, i.e. for each round t: $y_{<t}$ is accessible to the learning algorithm f_l.

- The number of experts $\mathcal{F}$ is finite or its $Ldim$ is finite.

There are also optimality results for cases where these conditions are only partly satisfied (e.g. the optimality results about intermittent prediction games or results for unboundedly growing or even infinite set of players in Schurz 2019, chpts.7-9).

This concludes our investigation of the logic of deceivability. Let us come to the applications now. We will start with traditional problems of epistemology and then go on with applications to the realm of social epistemology.

Part II

Optimisation in the Classical Epistemic Realm

Chapter 5

Induction and Hume's Problem

In this chapter the problem of induction is formulated and its emergence in the modern era is outlined. Afterwards, traditional and modern approaches are discussed and problems thereof are highlighted. Subsequently, more general learning theoretical impossibility results are presented which show that Hume's problem cannot be accounted for in terms of absolute learnability. Finally, it is outlined why the problem might be accounted for in terms of relative learnability and how the approach of meta-induction allows for justifying induction in this sense.

There are three major types of inference used in science and philosophy: deduction, induction, and abduction. Deductive inferences are characterised by their feature of truth preservation with certainty (or preservation of a ranking of truth values) in passing from the premisses to the conclusion. Induction, at least in the sense of *enumerative induction*, is characterised as non-deductive inference which has as a conclusion a generalised claim containing predicates that occur already in the premisses. Finally, abduction is formally characterised as a non-deductive inference with a conclusion containing also predicates other than that of the premisses.

A more general classification results from distinguishing only between deductive inferences (by their feature of truth preservation with certainty), and inductive inferences in the wide sense (see, e.g., Carnap 1952; and the discussion in Schurz 2019, sect.1.1). In this classification, the latter is an umbrella term for all non-deductive inferences, so also for inductive inferences as mentioned before (these are inductive inferences in the narrow sense). As we will see later on, so-called *creative abduction* is also about hypothesis or theory invention. Since philosophy of science is mainly about the context of justification of theories, but not about their utilisation and discovery, there is quite a big controversy whether abduction can be adequately dealt within philosophy of science, whether there is something like a "logic of abduction". For this reason, abduction is often not included in the umbrella term of induction in the sense of non-deductive inferences. Here we stick mainly to the former classification, however, sometimes we will also use the latter, more general one. Whenever we do so, we will make this clear by speaking of *induction in the wide* or *narrow sense*.

In this part of the book we will discuss the problem of optimisation with respect to these inferences: induction (this chapter and chapter 6), abduction (chapter 7), and deduction (chapter 8). Especially in our investigation of the problem of how to justify inductive inferences we will apply the results of the preceding part of the book.

In our investigation of the problem of induction, we first present the problem (section 5.1). Then we provide a short overview of traditional and modern approaches and indicate the main problems of these approaches (section 5.2). Afterwards, we link the problem of induction to the problem of absolute learnability, i.e. the idea that a prediction method has to be absolutely successful in the long run (section 5.3). Finally, we show that once one allows for justification via relative learnability in the sense of providing predictions which might fall apart from being successful, but which are optimal in the long run, one can also account for the problem of how to justify induction (section 5.4).

5.1 The Problem of Induction

Hume (1711-1776) wrote about the problem of induction, when inductive methods were already well established in science. One of the first modern physicists, Galileo Galilei (1564-1642), was already applying inductive methods when putting forward and testing his basic principle of relativity. And his contemporary Francis Bacon (1561-1626), the first "philosopher of the new physics" (see Hacking 2006, p.25), had already argued at length for the importance of inductive methods in science. He did so by discussing the *Novum Organum Scientiarum*, induction, in contrast to the *Organum Vetus* of Aristotle, deduction:

> "In ordinary logic almost all effort is concentrated on the syllogism. The logicians seem scarcely to have thought about induction. They pass it by with barely a mention, and hurry on to their formulae for disputation. But we reject proof by syllogism, because it [...] lets nature slip out of our hands. [...] For we regard induction as the form of demonstration which respects the senses, stays close to nature, fosters results and is almost involved in them itself." (see Bacon 1620/2000, p.16)

Also Isaac Newton's (1642-1726/27) *Principia Mathematica* were already widely accepted and its background methodology was gradually enriched in the several editions (1687, 1713, 1726) especially by methodological notes on the inductive method (see Feldbacher-Escamilla 2019). Induction was already well entrenched in the methodology of science, when Hume influentially put forward a problem of this method.

Well known is Hume's formulation of the problem of induction in the form of the following dilemma: Justification is either deductive or inductive. A principle of induction cannot be justified deductively, since it allows for inferences which are not truth preserving. It can be also not justified inductively, since this would amount to circular reasoning. Hence, we lack a justification for a principle of induction. Hume famously stated the *problem of induction* the first time in book I (*Of the Understanding*), part III (*Of knowledge and probability*), sect. VI (*Of the inference from the impression to the idea*) of his *A Treatise of Human Nature*, published 1738. There the above mentioned dilemma can be found in more or less explicit form as follows (see Hume 1738/1960, pp.86ff):

> *Ad: Justification is either deductive or inductive.*
> "Since it appears, that the transition from an impression present to the memory or senses to the idea of an object, which we call cause or effect, is founded on past experience, and on our remembrance of their *constant conjunction*, the next question is, whether experience produces the idea by means of the understanding or of the imagination; whether we are determin'd by reason to make the transition, or by a certain association and relation of perceptions. If reason determin'd us, it would proceed upon that principle, *that instances, of which we have had no experience, must resemble those, of which we have had experience, and that the course of nature continues always uniformly the same.*" (pp.88f)
>
> *Ad: There is no deductive justification of induction.*
> "[However], there can be no demonstrative arguments to prove, *that those instances, of which we have had no experience, resemble those, of which we have had experience*. We can at least conceive a change in the course of nature [... which is] a refutation of any pretended demonstration[.]" (p.89)
>
> *Ad: There is no inductive justification of induction.*
> "[On the other hand] 'Tis therefore necessary, that in all probable reasonings there be something present to the mind, either seen or remember'd; and that from this we infer something connected with it, which is not seen nor remember'd. The only connexion or relation of objects, which can lead us beyond the immediate impressions of our memory and senses, is that of cause and effect; [...] The idea of cause and effect is deriv'd

> from *experience*, which informs us, that such particular objects, in all past instances, have been constantly conjoin'd with each other: And as an object similar to one of these is suppos'd to be immediately present in its impression, we thence presume on the existence of one similar to its usual attendant. According to this account of things, which is, I think, in every point unquestionable, probability is founded on the presumption of a resemblance betwixt those objects, of which we have had experience, and those, of which we have had none; and therefore 'tis impossible this presumption can arise from probability. The same principle cannot be both the cause and effect of another;" (pp.89f)

Disappointed with the reception of his *Treatise*, Hume published *An Enquiry Concerning Human Understanding* in which he elaborated his thoughts in a shorter and more polemical way ten years later (1748). There his thoughts on induction appear in section IV (*Sceptical doubts concerning the operations of the understanding*). We can explicate his argument as follows (see Hume 1748/2007, pp.24ff, 33–36):

1. "All reasonings may be divided into two kinds, namely demonstrative reasoning, or that concerning relations of ideas, and moral reasoning, or that concerning matter of fact and existence." (p.25, 35)
   Modern paraphrase: All justification stems from either analytic or synthetic reasoning.
   Schematically: $Jx \leftrightarrow (Ax \vee Sx)$

2. "When it is asked, *What is the nature of all our reasonings concerning matter of fact?* the proper answer seems to be, that they are founded on the relation of cause and effect. When again it is asked, *What is the foundation of all our reasonings and conclusions concerning that relation?* it may be replied in one word, *Experience*." (p.23, 32)
   Modern paraphrase: Synthetic reasoning (might be causal reasoning, but) is ultimately based on experience alone.
   Schematically: $(Sx \rightarrow Cx)$ & $(Cx \rightarrow Ex)$

3. "It must certainly be allowed, that nature has kept us at a great distance from all her secrets, and has afforded us only the knowledge of a few superficial qualities of objects; while she conceals from us those powers and principles, on which the influence of these objects entirely depends. [...] As to past *Experience*, it can be allowed to give *direct* and *certain* information of those precise objects only, and that precise period of time, which fell under its cognizance: But why this experience should be extended to future times, and to other objects, which for aught we know, may be only in appearance similar; this is

the main question on which I would insist. [...] It must be acknowledged, that there is here a consequence drawn by the mind; that there is a certain step taken; a process of thought, and an inference, which wants to be explained. These two propositions are far from being the same, *I have found that such an object has always been attended with such an effect*, and *I foresee, that other objects, which are, in appearance, similar, will be attended with similar effects*." (pp.24f, 33f)
Modern paraphrase: The principle of induction is not based on experience alone.
Schematically: $\neg E p_i$

4. "That there are no demonstrative arguments in the case, seems evident; since it implies no contradiction, that the course of nature may change, and that an object, seemingly like those which we have experienced, may be attended with different or contrary effects. [...] Whatever is intelligible, and can be distinctly conceived, implies no contradiction, and can never be proved false by any demonstrative argument or abstract reasoning *à priori*." (p.25, 35)
Modern paraphrase: The principle of induction is not analytic.
Schematically: $\neg A p_i$

5. Hence: "If we be, therefore, engaged by arguments to put trust in past experience, and make it the standard of our future judgment, these arguments must be probable only, or such as regard matter of fact and real existence." (p.26, 35)
Modern paraphrase: Hence, if the principle of induction is justified, then by synthetic reasoning.
Schematically: $J p_i \rightarrow S p_i$

6. Hence: "But that there is no argument of this kind, must appear, if our explication of that species of reasoning be admitted as solid and satisfactory. We have said, that all arguments concerning existence are founded on the relation of cause and effect; that our knowledge of that relation is derived entirely from experience; and that all our experimental conclusions proceed upon the supposition, that the future will be conformable to the past. To endeavour, therefore, the proof of this last supposition by probable arguments, or arguments regarding existence, must be evidently going in a circle, and taking that for granted, which is the very point in question." (p.26, 35f)
Modern paraphrase: Hence, the principle of induction is not justified.
Schematically: $\neg J p_i$

How influential Hume's sceptical argument against induction was, is a matter of historical dispute. Hacking (2006), e.g., argues that it was necessary that the underlying notions of *probability*, *induction*, and *statistical inference* allowed for a distinction between opinion and knowledge as a matter of degree, and that before Hume's *Treatise* this was not the case and so his argument played a central role already quite early on. Laudan (1981), in contrast, argues that:

> "It is one of the wilder travesties of our age that we have allowed the myth to develop that 19th-century philosophers of science were as preoccupied with Hume as we are. As far as I have been able to determine, none of the classic figures of 19th-century methodology—neither Comte, Herschel, Whewell, Bernard, Mill, Jevons, nor Peirce—regarded Hume's arguments about induction as much more than the musings of an historian. This claim is borne out by the fact that in Peirce's thirty-two papers on induction and scientific method—papers teeming with historical references—there is only one reference to Hume; and that is not in connection with the problem of induction but with the problem of miracles." (Laudan 1981, p.240)

And he adds, tightening his claim: "Hume's avoidance of [... the notion of] induction [as we understand it nowadays] was probably related to his almost unparalleled ignorance of the science of his time. [... In footnote 38:] Indeed, it is difficult to find a major philosopher between Socrates and G. E. Moore who knew less than Hume about the science of his time" (see Laudan 1981, pp.83f).

There could not be any sharper conflict between two positions on the reception of Hume's thoughts. These two positions span from necessary influence due to the underlying conceptual development to almost full negligence. However, it seems that Reichenbach (1938) proposed already a middle ground which allows for an explanation of both, negligence (by classical empiricists) and influence (of logical empiricists due to a new formalism):

> "If inductive inference can teach us something new, in opposition to deductive inference, this is because it is not a tautology. [...] It was David Hume who first attacked the principle from this side; he pointed out that the apparent constraint of the inductive inference, although submitted to by everybody, could not be justified. [...] We may summarize his objections in two statements: We have no logical demonstration for the validity of inductive inference. There is no demonstration a posteriori for the inductive inference; any such demonstration would presuppose the very principle which it is to demonstrate. [...] In spite of the deep impression Hume's discovery made on his contemporaries, its relevance was not sufficiently noticed in the subsequent intellectual development. [...] It is astonishing to see how clear-minded logicians, like John Stuart Mill, or Whewell,

> or Boole, or Venn, in writing about the problem of induction, disregarded the bearing of Hume's objections; they did not realize that any logic of science remains a failure so long as we have no theory of induction which is not exposed to Hume's criticism. [...] It remains incomprehensible that their empiricist principles did not lead them to attribute a higher weight to Hume's criticism. It has been with the rise of the formalistic interpretation of logic in the last few decades that the full weight of Hume's objections has been once more realized. [...] Hume's criticism was the heaviest blow against empiricism; if we do not want to dupe our consciousness of this by means of the narcotic drug of aprioristic rationalism, or the soporific of skepticism, we must find a defense for the inductive inference which holds as well as does the formalistic justification of deductive logic." (see Reichenbach 1938, pp.341f, p.347)

In support of Reichenbach's argument we point to the fact that one of the first to highlight (and appreciate) Hume's scepticism regarding induction was John Maynard Keynes who counts as one of the forerunners of those philosophers of science who provided a new formalism for inductive reasoning (the programme of *logical probabilities*):

> "Between Bacon and Mill came Hume. Hume's sceptical criticisms are usually associated with causality; but argument by induction—inference from past particulars to future generalisations—was the real object of his attack. Hume showed, not that inductive methods were false, but that their validity had never been established and that all possible lines of proof seemed equally unpromising. The full force of Hume's attack and the nature of the difficulties which it brought to light were never appreciated by Mill, and he makes no adequate attempt to deal with them." (Keynes 1921, pp.312f)

Regardless of the exact influence of Hume's thought on his contemporaries, it is interesting to note that in fact the term 'induction' does not at all appear in Hume's argument, nor (almost) anywhere in the *Treatise* or the *Enquiry* (Vickers 2010, sect.2). Rather, Hume speaks mainly about inferences concerning causal connections. However, it seems that in his critique and discussion of such connections he clearly had in mind a principle of enumerative induction: If all objects of one kind (e.g. objects experienced in the past) have a property (e.g. that a cause is conjoined with the effect), then also objects of another kind (e.g. objects to be experienced in the future) are supposed to have that property. Since enumerative induction is about inferences to generalisations, asking for a justification of this problem is to ask for a justification of a generalising inference (see Vickers 2010, sect.1).

However, if one construes the notion of an *inductive inference* more broadly and considers, e.g., as an inductive inference all non-deductive (and non-abductive) inferences (see, e.g., Carnap 1952), then the justification problem can be also reframed as a *characterisation* or *demarcation problem*, namely the problem of how to characterise or demarcate *good* inductions in contrast to *bad* ones (see Vickers 2010, sect.1).

In the following sections of this and the next chapter we are after the problem of induction in this more general sense: Can one provide epistemic justification for any non-deductive (and non-abductive) method?

5.2 Traditional and Modern Approaches

In this section we are going to discuss traditional approaches to the problem of induction. Since the literature on this topic is enormous, we aim only at providing a brief sketch for an approach of each of the main positions discussed in our investigation of the problem of epistemic justification (chapter 1). Recall, the approaches to epistemic justification we considered there were: *foundationalism*, *coherentism*, *infinitism*, and *naturalised epistemology*. We will discuss a foundationalist, coherentist, infinitist, and a naturalised solution to the problem of induction in this section (in reverse ordering). We will also discuss more modern approaches, namely the falsificationist approach of Popper as well as the approach of inductive logic. In the subsequent section we provide a formal argument for being *sceptical* about a strict solution to the problem (a solution in terms of absolute learnability). Finally, in section 5.4 we describe the solution put forward by *epistemic engineering*.

A Naturalised Approach

Let us begin with a *naturalised* approach to the problem of induction: Hume himself offers an approach to the problem of induction which might be assigned to this programme. E.g., in section V (*Sceptical solution of these doubts*) of the *Enquiry*, he argues:

1. "[Suppose one has] observed similar objects or events to be constantly conjoined together; what is the consequence of this experience?" (p.31, 42)

2. "He immediately infers the existence of one object from the appearance of the other." (p.31, 42)

3. "Though he should be convinced, that his understanding has no part in the operation, he would nevertheless continue in the same course of thinking. There is some other principle, which determines him to form such a conclusion." (p.32, 42)

4. "This principle is *Custom* or *Habit*. For wherever the repetition of any particular act or operation produces a propensity to renew the same act or operation, without being impelled by any reasoning or process of the understanding; we always say, that this propensity is the effect of *Custom*. By employing that word, we pretend not to have given the ultimate reason of such a propensity. We only point out a principle of human nature, which is universally acknowledged, and which is well known by its effects. Perhaps, we can push our enquiries no farther, or pretend to give the cause of this cause; but must rest contented with it as the ultimate principle[.]" (p.32, 43)

So, according to Hume we are accustomed to make inductive inferences and that is all about it. This strategy is sometimes also called "*explain where you can't justify*" (see Howson 2000, p.21). However, the achievements of such a strategy are too modest in order to be satisfying for the bulk of epistemologists. Similarly, as in general a criminal cannot justify her behaviour by simply providing a detailed description of how she committed the crime, we do not accept an explanation of de facto inductive reasoning by custom as a justification. As we have seen in our discussion of naturalised epistemology in section 1.4, one needs to account for the normative part of justification by more than providing descriptive principles. In the case of *custom*, we would need justification for a principle which states that custom is a necessary or an optimal means to achieve our epistemic goals like truth or accurate predictions. However, a justification of such a principle, e.g. by reference to past success of the methods we are accustomed with, is exposed to the problem of induction again.

An Infinitist Approach

Let us come to another approach which might be subordinated to *infinitism*. This approach tackles the second horn of Hume's dilemma and claims to argue for induction by help of induction in a non-circular way. A proponent of this approach is, e.g., John St. Mill who tried to justify induction by reference to the uniformity of nature (see Chapter iii: Of the Ground of Induction Mill 1843/1974, pp.1106-1110). A modern form of such a non-circular approach is provided by M. Black (1954, Inductive Support of Inductive Rules). Our description of such an infinitist account is due to (see Skyrms 2000, sect.III.3). The idea is as follows: Assume we want to infer by help of induction that the next observed r*A*ven is *B*lack ($Aa_{1_n} \rightarrow Ba_{1_n}$) on the basis that all up to now observed ravens were black ($Aa_{1_1}\&Ba_{1_1},\ldots,Aa_{1_{n-1}}\&Ba_{1_{n-1}}$). We can do so by help of our experience as well as a level 1 principle of induction about the individuals. The argument at level 1 is as follows:

$$a_{2_1}\left\{\begin{array}{l} Aa_{1_1}\&Ba_{1_1},\ldots,Aa_{1_{n-1}}\&Ba_{1_{n-1}}\\ Aa_{1_1}\&Ba_{1_1},\ldots,Aa_{1_{n-1}}\&Ba_{1_{n-1}}\mid\!\sim_1\ Aa_{1_n}\rightarrow Ba_{1_n}\\ \text{Hence: } Aa_{1_n}\rightarrow Ba_{1_n}\end{array}\right.$$

Here '$\mid\!\sim_1$' stands for a level 1 inference and 'a_{1_i}' stands for ordinary objects like ravens, shoes etc. Now, how to justify this principle of induction? One can do so by providing a level 2 argument with a level 2 principle of induction about level 1 inferences: One argues from past experience about the Truth-conduciveness of level 1 Inductions ($I_1a_{2_1}\&Ta_{2_1},\ldots,I_1a_{2_{m-1}}\&Ta_{2_{m-1}}$) and a level 2 principle of induction about level 1 induction as follows:

$$a_{3_1}\left\{\begin{array}{l} I_1a_{2_1}\&Ta_{2_1},\ldots,I_1a_{2_{m-1}}\&Ta_{2_{m-1}}\\ I_1a_{2_1}\&Ta_{2_1},\ldots,I_1a_{2_{m-1}}\&Ta_{2_{m-1}}\mid\!\sim_2\ I_1a_{2_m}\rightarrow Ta_{2_m}\\ \text{Hence: } I_1a_{2_m}\rightarrow Ta_{2_m}\end{array}\right.$$

Note that 'a_{2_i}' stands not for ordinary objects like ravens, shoes etc., but for inferences or arguments as, e.g., labelled by 'a_{2_1}' above. E.g. '$I_1a_{2_1}\&Ta_{2_1}$' is to be interpreted as: the above argument (a_{2_1}) is an inductive inference and was successful or truth-conducive. Now, how to argue for the level 2 principle of induction? One can do so by providing a level 3 argument with a level 3 principle of induction about the Truth-conduciveness of level 2 Inductions. For the level 3 principle of induction one can argue likewise. In general, one can argue for a level n principle of induction by help of an $n+1$ principle of induction and information about level n principle's past success:

$$\begin{array}{l} I_na_{n+1_1}\&Ta_{n+1_1},\ldots,I_na_{n+1_{l-1}}\&Ta_{n+1_{l-1}}\\ I_na_{n+1_1}\&Ta_{n+1_1},\ldots,I_na_{n+1_{l-1}}\&Ta_{n+1_{l-1}}\mid\!\sim_{n+1}\ I_na_{n+1_l}\rightarrow Ta_{n+1_l}\\ \text{Hence: } I_na_{n+1_l}\rightarrow Ta_{n+1_l}\end{array}$$

And so on, in principle ad infinitum. Note that the reasoning is not circular, because one always provides different evidence for different principles of induction. I_i, T-statements about the past and present are experienced. And the justification of $\mid\!\sim_1$ is provided by help of such experience and $\mid\!\sim_2$, that of $\mid\!\sim_2$ by help of such experience and $\mid\!\sim_3$, and more generally that of $\mid\!\sim_n$ by help of $\mid\!\sim_{n+1}$. The schema of this infinitist inductive reasoning for induction is depicted in figure 5.1.

Note that suitable information about the method's past success is needed. E.g., if we get to know that level 1 inductions do not work properly in the sense that past level 1 inductive inferences were not truth conducive, then also no level 2 inductive inference can be performed, since such an inference is licensed only on the basis of the past truth conduciveness of level

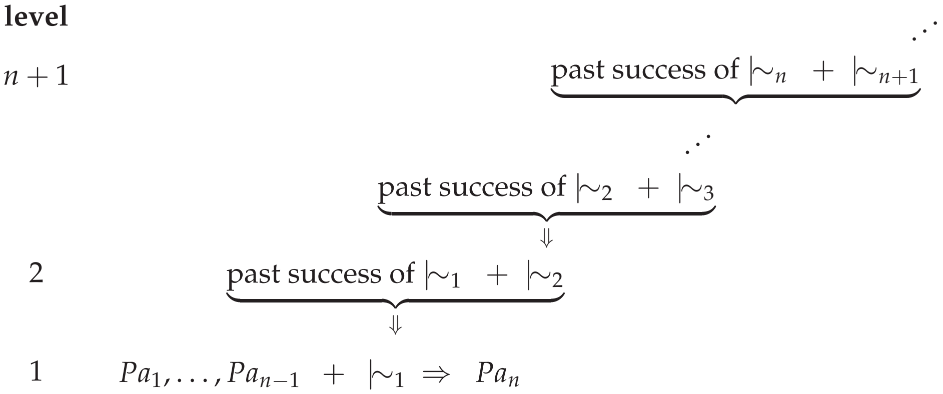
level
$n+1$
past success of $|\!\!\sim_n$ + $|\!\!\sim_{n+1}$
past success of $|\!\!\sim_2$ + $|\!\!\sim_3$
2
past success of $|\!\!\sim_1$ + $|\!\!\sim_2$
1
$Pa_1, \ldots, Pa_{n-1}$ + $|\!\!\sim_1$ ⇒ Pa_n

Figure 5.1: Schema of an infinitist inductive justification of induction by arguing with inductive principles on different levels: On level 1 object properties P are *transferred* from the past and presence to the future. On level 2 success properties from past and present level 1 inferences are *transferred* to the future. And so on ad infinitum. (see table III.2 of Skyrms 2000, p.38)

1 inferences. In general, once one accepts an infinitist notion of justification (J), it seems that one can provide a non-circular inductive justification of induction. Is infinitism a viable solution to the problem of induction? As the following argument shows, at least this version of an infinitist justification of induction fails.

The reason for this is that this schema of infinitism allows not only for justifying induction, but also anti-induction or counterinduction or counter-conduction (see W. C. Salmon 1957; and Skyrms 2000, pp.41ff): Let us begin with an anti-inductive argument at level 1 as follows:

$$a_{2_1} \begin{cases} Pa_{1_1}, \ldots, Pa_{1_{n-1}} \\ Pa_{1_1}, \ldots, Pa_{1_{n-1}} \; ||\!\!\sim_1 \; \neg Pa_{1_n} \\ \text{Hence: } \neg Pa_{1_n} \end{cases}$$

Now, how can we argue for $||\!\!\sim_1$? Clearly, we cannot do so by help of a level 2 inductive principle $|\!\!\sim_2$, because $||\!\!\sim_1$ did not work well in past (either $|\!\!\sim_1$ or $||\!\!\sim_1$ worked well in past, but not both). However, we can provide a level 2 anti-inductive principle $||\!\!\sim_2$ stating that what was successful in past, wont be successful in the future or what was not successful in past, will be successful in the future. Arguing from the $\neg T$ruth-conduciveness of level 1 aNti-inductions, a level 2 anti-inductive principle allows for the following argument, justifying the anti-inductive inference on level 1:

$$a_{3_1} \begin{cases} N_1 a_{2_1} \& \neg Ta_{2_1}, \ldots, N_1 a_{2_{m-1}} \& \neg Ta_{2_{m-1}} \\ N_1 a_{2_1} \& \neg Ta_{2_1}, \ldots, N_1 a_{2_{m-1}} \& \neg Ta_{2_{m-1}} \; ||\!\!\sim_2 \; N_1 a_{2_m} \rightarrow Ta_{2_m} \\ \text{Hence: } N_1 a_{2_m} \rightarrow Ta_{2_m} \end{cases}$$

In the same way, more generally, an anti-inductive principle at level n can be justified by help of the same experience as before (past success of $|\!\sim_n$ iff past failure of $||\!\sim_n$ and vice versa) and an anti-inductive principle at level $n+1$. The schema of this infinitist anti-inductive reasoning for anti-induction is depicted in figure 5.2.

$$\begin{array}{ll}
\textbf{level} & \hfill \iddots \\
n+1 & \hfill \underbrace{\text{past failure of } ||\!\sim_n \;+\; ||\!\sim_{n+1}} \\
 & \hfill \iddots \qquad\qquad\quad \\
 & \hfill \underbrace{\text{past failure of } ||\!\sim_2 \;+\; ||\!\sim_3} \qquad\qquad\qquad \\
 & \hfill \Downarrow \qquad\qquad\qquad\qquad\qquad\quad \\
2 & \qquad\qquad \underbrace{\text{past failure of } ||\!\sim_1 \;+\; ||\!\sim_2} \\
 & \qquad\qquad\qquad\qquad \Downarrow \\
1 & Pa_1, \ldots, Pa_{n-1} \;+\; ||\!\sim_1 \;\Rightarrow\; \neg Pa_n
\end{array}$$

Figure 5.2: Schema of an infinitist anti-inductive justification of anti-induction by arguing with anti-inductive principles on different levels: On level 1 object properties P are *inversed* from the past and presence to the future. On level 2 failure properties from past and present level 1 inferences are *inversed* for the future. And so on ad infinitum. (see table III.3 of Skyrms 2000, p.43)

Again, once one accepts an infinitist notion of justification (J), it seems that one can provide also a non-circular anti-inductive justification of anti-induction. So, this infinitist approach to the problem of induction allows for justifying principles or inference methods with contradicting consequences or conclusions. Note that in this case, an infinitist could, e.g., argue likewise as a coherentist does when facing the problem of licensing contradicting inferences or systems: She might just state that not all infinitely justify*able* principles and inferences *are* justified, but only logically compatible ones. So, either induction or anti-induction is justified by this infinitist approach, but not both. However, then one still is in need of an argument for choosing among them. In this sense the notion of *justification* is still epistemically underdetermined (see our discussion at the end of section 1.2).

A Coherentist Approach

This brings us to the next approach to the problem of induction, namely coherentism. Recall, according to Hume, an inductive justification of induction "must be evidently going in a circle" (see Hume 1748/2007, p.26, 35). Infinitism suggested to consider a different notion of *justification* and tried to overcome the problem by an infinite hierarchy of reasoning, however failed as we have outlined above. But what about accepting a notion of *justification* (J) which allows for circular reasoning. One idea of such a coherentist approach to the problem of justifying induction is that an inductive principle might more or less directly provide epistemic support for itself. Now, if we take a very fundamental stance of coherentism according to which the only relevant criterion for justification is logical consistency, then any conservative expansion of logical inferences would be licensed. Note that an inductive principle $Pa_1, \ldots, Pa_{n-1} \mid\!\sim Pa_n$ as well as an anti-inductive principle $Pa_1, \ldots, Pa_{n-1} \mid\mid\!\sim \neg Pa_n$ allow for conservatively expanding classical logic. Hence, according to this criterion both, induction as well as anti-induction, would be justified (see Schurz 2019, sect.3.3). So, such an approach ends up with the same problem as we discussed before for the infinitist approach to the problem of induction. Also putting forward coherence standards with coherence constraints on success fail. If we assume, e.g., that an inference rule is justified, if it is guaranteed to be successful "according to its own inferences regarding success", then such a success coherent approach to justification licenses induction (success coherentism demands that the assumption of successful inferences needs to be coherent with allowing for replacing the property of objects P by the success property of principles and inferences):

1. Inductive inferences have been successful in past.
2. Therefore, by the rule of induction, inductive inferences will be successful in the future.
3. Therefore, it is (internally) coherent to assume that inductive inferences will be successful.

Clearly, anti-induction is not success coherent according to *inductive* inferences regarding success:

1. Anti-inductive inferences have failed in past.
2. Therefore, by the rule of *induction*, anti-inductive inferences will fail in the future.
3. Therefore, it is *not* (externally based on induction) coherent to assume that anti-inductive inferences will be successful.

However, also induction is not success coherent according to *anti-inductive* inferences regarding success:

1. Inductive inferences have been successful in past.
2. Therefore, by the rule of *anti-induction*, inductive inferences will fail in the future.
3. Therefore, it is *not* (externally based on anti-induction) to assume that inductive inferences will be successful.

Regarding "its own inferential standards concerning success", a success coherent approach to justification licences also anti-induction:

1. Anti-inductive inferences have failed in past.
2. Therefore, by the rule of anti-induction, anti-inductive inferences will be successful in the future.
3. Therefore, it is (internally) coherent to assume that anti-inductive inferences will be successful.

So, the same problem of infinitism shows up not only for naïve coherentism in the sense of licensing any inference which is logically coherent, but also for success coherentism licensing any inference which is logically coherent and coherent with the assumption of its own success. Below we will see that also a more sophisticated form of coherentism (namely probabilism) fails with respect to justifying induction.

A Foundationalist Approach

Let us also very briefly sketch a foundationalist approach to the problem of induction. We will only rush through it, since such an approach seems to trivialise the problem and faces the general problems we described already while discussing foundationalism in the context of the general problem of epistemic justification in section 1.1. According to foundationalism, the notion of *justification* (J) allows for some propositions to be justified, without there being any reason provided for them. We have described such propositions by help of the foundationalist principle (F) as an *epistemic basis* B. The idea is that all propositions are justified which are either in B itself or which can be deduced from B (i.e. for which elements of B serve as reasons). Now, a simple foundationalist approach to the problem of induction might consider a/the principle of induction to be an element of B, i.e. not in need of any reason. In this way the problem of induction vanishes for trivial reasons, since principle (F) allows for justification of the elements of B without any further reason except them being in B. However, why then not alternatively considering a/the principle of anti-induction to be an element of B? Clearly, the problem of the choice of B shows up again. In our discussion of foundationalism we have also seen that there are two approaches relevant for such a choice: internalism and externalism. Regarding an internalist account to the problem of choosing B we have seen in section 1.1 that a regress argument can be triggered by asking for the truth conduciveness of the choice. For this reason also an internalist foundationalist approach to the problem of induction fails. According to externalism, on the other hand, what counts for such a choice is only whether our choice of B leads to reliable inferences or not. There is no need for an epistemic agent to be actually aware of this or not. This means that choosing B with a principle of induction allows for justification, if induction is *de facto* a reliable inference method, and choosing B with a principle of anti-induction allows for justification, if anti-induction is *de facto* a reliable inference method. However, this means that the notion of *justification* is accessible to us only if we take in a *God's eye view*, something we "do" when we speak about truth, but something we would not expect for using the notion of *justification* as serving as an intermediary between *truth* and *belief*. Hence, also an externalist foundationalist approach to the problem of induction fails.

Popper's Falsificationist Approach

Finally, we also want to hint at two more modern approaches to Hume's problem which do not directly fit into the discussed branches of theories of epistemic justification. First, and very briefly, there is Popper's falsificationist approach. For Popper, Hume's problem was key for shifting the paradigm of scientific methodology from verification and confirmation to falsification and corroboration. He suggested to consider any falsifiable theory as scientific and to "redefine" the task of (philosophy of) science to try to falsify theories instead of confirming them. Justification (J) is, so to say, granted per default for any falsifiable theory, and withdrawn, in case a theory is in fact falsified. By applying such a methodology only deductive inferences are needed and Hume's problem vanishes simply by marginalising the role of induction:

> "The best we can say of a hypothesis is that up to now it has been able to show its worth, and that it has been more successful than other hypotheses although, in principle, it can never be justified, verified, or even shown to be probable. This appraisal of the hypothesis relies solely upon *deductive* consequences (predictions) which may be drawn from the hypothesis. *There is no need even to mention induction*." (Popper 2002b, p.317)

Although Popper's methodological shift was very influential and is *de facto* applied in form of *null hypothesis significance tests* (see Sprenger 2016, sect.1 and 6), it is not considered to cover adequately and fully the whole range of scientific practice. On the contrary, scientists often speak of *verification* and *confirmation* and as we will indicate now, also most of the philosophy of science literature on theory assessment is about confirmation. For this reason Hume's problem, although contested, still stands as it is.

The Approach of Inductive Logic and Confirmation Theory

This brings us to the second modern account, namely the approach of *inductive logic* (for an overview see Sprenger 2016). Inductive logic studies the notion of and measures for confirmation. This theory of confirmation covers a bulk of philosophy of science studies of the 20th century and at its core it studies principles which more or less circumvent Hume's problem. Early proponents of such theories of confirmation were Carl G. Hempel who had a qualitative approach to confirmation (see Hempel 1945a,b) and Carnap who was one of the first to provide a quantitative approach (see Carnap 1950/1962, 1952). According to Carnap, Hume's problem concerns the task of justifying inductive inferences, in particular their conclusions, i.e. hypotheses H, on the basis of some evidence E. In contrast to this the problem of inductive logic is to determine a measure for confirmation of some hypothesis H by the evidence E. The idea of Carnap was to define such a measure by help of logical or combinatorial principles—this is the so-called programme of *logical probabilities*. In providing such a measure one simply avoids the qualitative question of justification. Here is how Carnap describes the aim of inductive logic:

> "It seems to me that the view of almost all writers on induction in the past and including the great majority of contemporary writers, contains one basic mistake. They regard inductive reasoning as an *inference* leading from some known propositions, called the premisses or evidence, to a new proposition, called the conclusion, usually a law or a singular prediction. From this point of view the result of any particular inductive reasoning is the *acceptance* of a new proposition[. ...] This seems to me wrong. On the basis of this view it would be impossible to refute Hume's dictum that there are no rational reasons for induction. [...] I would think instead that inductive reasoning about a proposition should lead, not to acceptance or rejection, but to the assignment of a number to the proposition, viz., its [degree of confirmation]. This difference may perhaps appear slight; in fact, however, it is essential. If, in accordance with the customary view, we accept the prediction, then Hume is certainly right in protesting that we have no rational reason for doing so, since, as everybody will agree, it is still possible that [our prediction is wrong]. If, on the other hand, we adopt the new view of the nature of inductive reasoning, then the situation is quite different. In this case Input does not assert the hypothesis H in question, e.g., the prediction [rather its degree of confirmation. ... Note that this suffices for induction to] fulfil its purpose of guiding our practical decisions [... since] for the determination of a rational decision neither the acceptance of H nor knowledge of the objective probability of H is needed." (see Carnap 1966, p.317f)

So, the idea of Carnap's approach can be summarised as follows: Hume thought about induction in qualitative terms stating that the inference of a hypothesis H from some evidence E either is justified or not. Classical qualitative decision theory then states that H is accepted iff its inference from E is justified or admissible and rejected otherwise. Alternatives of H play no role here. And whether we accept H or not hinges completely on whether the inference from E to H is justified. However, if one switches to a quantitative consideration and modern decision theory, then one seems to get fully rid of the question of justification: One considers a collection of alternative hypotheses $H, H', H'', \ldots$ and provides for each hypothesis its degree of confirmation given the evidence E: $conf(H, E)$, $conf(H', E)$, $conf(H'', E)$, ... Then, decision theory demands to opt for that hypothesis whose degree of confirmation maximises the underlying utilities (in this simplified picture we consider $conf$ to be the so-called *absolute measure of confirmation* which consists in the conditional probability of H given E). According to Carnap, no problem of justification shows up, simply because the decision theoretic framework is justified via optimality considerations, and the measure of confirmation is supposed to be a purely logical or combinatorial measure. Regarding the justification of the decision theoretic framework we have discussed already in section 1.4 that it stems from optimality considerations and that these are commonly accepted also in other areas where normativity considerations are relevant, as, e.g., in ethics.

But how about the justification of the measure for the degree of confirmation? Whether one has to opt for Pa_n (H) or $\neg Pa_n$ (H') given evidence $Pa_1, \ldots, Pa_{n-1}$ (E), clearly depends on whether $conf(H, E) > conf(H', E)$ or not (given equal underlying utilities), i.e.: An (enumerative) inductive inference is justified compared to an anti-inductive inference only, if:

$$conf(Pa_n, Pa_1 \& \cdots \& Pa_{n-1}) > conf(\neg Pa_n, Pa_1 \& \cdots \& Pa_{n-1})$$

Now, Carnap's idea was to argue for a measure $conf$ which satisfies this condition (also called a *singular predictive inference* (see Carnap 1966, §110C, pp.567f)) in a similar way as one might argue for the validity of deductive inferences:

> "We shall see that a statement of deductive logic like 'e L-implies h' means the entire range of e is included in that of h, while a statement of inductive logic like '$\mathfrak{c}(h, e) = 3/4$' means three-fourths of the range of e is included in that of h." (Carnap 1950/1962, p.202)

For illustrative purposes we outline this approach in a nutshell. Considering only propositional logic with $p_1, \ldots, p_n$ propositional variables, we can define a *state description* as the conjunction of $\pm p_1 \& \cdots \& \pm p_n$ (where $\pm$ means that the respective propositional variable is either negated or not negated). The *range* of some p is the set of those state descriptions where

	Pa_1	Pa_2	State Description
1	0	0	$\neg Pa_1 \& \neg Pa_2$
2	0	1	$\neg Pa_1 \& Pa_2$
3	1	0	$Pa_1 \& \neg Pa_2$
4	1	1	$Pa_1 \& Pa_2$

Table 5.1: Example of the state descriptions for a monadic first order language $\mathcal{L}^{2,1}$ with two individual constants and 1 monadic predicate.

p is true (not negated). One can think of the range of a some p also as the set of disjunctive elements of p's disjunctive normal form. E.g., if $n = 2$, then the disjunctive normal form of p_1 is $p_1 \& p_2 \vee p_1 \& \neg p_2$, whereas the disjunctive normal form of $p_1 \& p_2$ is $p_1 \& p_2$ itself. So, the range of p_1 is $\{p_1 \& p_2, p_1 \& \neg p_2\}$, whereas the range of $p_1 \& p_2$ is $\{p_1 \& p_2\}$. That p_1 is a deductive consequence of $p_1 \& p_2$ means that the entire range of $p_1 \& p_2$ is included in that of p_1, which is true. Clearly, $p_1 \& p_2$ is not a deductive consequence of p_1 since the range of the latter is not entirely included in that of the former. However, it is partially included. As stated above, Carnap's idea was to consider the degree of inclusion also as the degree of confirmation. Since 1 out of 2 elements of p_1's range is contained in that of $p_1 \& p_2$, the degree of confirmation of $p_1 \& p_2$ by p_1 is 1/2. Our simple inductive logic for finitely many propositional variables can be straightforwardly expanded to a monadic first-order language (without quantifiers) with n individual constants and m monadic predicates: $\mathcal{L}^{n,m}$. E.g., if $n = 2$ and $m = 1$ we get the four state descriptions provided in table 5.1 (see Carnap 1950/1962, pp.106f). This framework allows us already to formulate the question of the validity of enumerative induction in comparison to anti-induction. Given a monadic first-order language $\mathcal{L}^{n,1}$: To which degree is the range of $Pa_1 \& \cdots \& Pa_{n-1}$ included in that of Pa_n, and to which degree is it included in that of $\neg Pa_n$. Furthermore, which one of both is higher? Now, the answer is simple: Both, Pa_n and $\neg Pa_n$, are satisfied by equally many state descriptions (2^{n-1}). Furthermore, there are equally many state descriptions in the range of $Pa_1 \& \cdots \& Pa_{n-1}$ which satisfy Pa_n as there are which satisfy $\neg Pa_n$ (namely for each 1). Hence, $conf(Pa_n, Pa_1 \& \cdots \& Pa_{n-1}) = conf(\neg Pa_n, Pa_1 \& \cdots \& Pa_{n-1})$. So, according to this simple idea, inductive inferences are "logically" not better off than anti-inductive ones (that the no free lunch theorem is a generalisation of this fact is shown in Schurz 2017). For this reason, Carnap introduced a new parameter, λ, which should be an inverse measure for the speed of learning from experience (see Carnap 1952; and the outline in Carnap 1950/1962, §110C, p.568; also Carnap 1959, p.218): Let s_i be the number of H-instances in the evidence E: $s_i = |\{a_i : E \vdash H[a_n/a_i]\}|$; let s be the sample size (i.e. the number of individual constants in E), and let κ be the number of so-called *Q-predicates* of $\mathcal{L}^{n,m}$ (where a Q-predicate is a "maximally consistent predicate" definable in $\mathcal{L}^{n,m}$ and has the form $\pm P_1x\& \cdots \& \pm P_mx$; so $\kappa = 2^m$). Furthermore, let w be the width of the hypothesis H, which is defined as the number of Q-predicates whose disjunction is equivalent with H. Then he defines a measure of confirmation as:

$$conf^{\lambda}(H,E) = conf^{\lambda}(H[a_n], E[a_1, \ldots, a_{n-1}]) = \frac{s_i + \frac{\lambda \cdot w}{\kappa}}{s + \lambda}$$

If we ignore for a moment λ, κ and w and if we describe permutability of the individuals within an expression as the structure of the expression, then $conf$ is no longer about the degree of inclusion of E's range in that of H, but about the inclusion of E's structure in that of H. In our simple singular predictive inference we consider just one predicate, hence $\kappa = 2$. The learning parameter λ can take on any value in $[0, \infty)$: Learning can be *super* fast or *super* slow in the sense of impossible. This allows for a whole spectrum, a continuum of inductive methods. Well known are the following specifications (see Carnap 1950/1962, §110C, p.568; and Sprenger 2016, sect.3):

- $\lambda = 0$: *Straight rule*, transferring the frequency of the observed sample to the unobserved case: $conf^0(H,E) = \frac{s_i}{s}$
- $\lambda = 2$: *Laplace's rule of succession*, looking at an experiment for which both success and failure are possible, and estimates as if we had observed one success and one failure, so predicting like the straight rule where one performed $s + 2$ experiments and found $s_i + 1$ positive instances (here the relative width w/κ is $1/2$): $conf^2(H,E) = \frac{s_i+1}{s+2}$
- $\lambda \to \infty$: *Inductive scepticism*, according to which we cannot learn anything from past experience about the future: $conf^{\infty}(H,E) = 0$

Now, Carnap's programme of *logical probability* is generally considered to be a degenerative research programme, inasmuch as generalisations for quantified first-order logic and infinite domains are tricky and in need of several parametrisations. Regardless of this, we see already with the *free* λ parameter that whether inductive inferences are justified or not and if so, to which degree, depends on the choice of λ. For small λ, $conf^{\lambda}(Pa_n, Pa_1\& \cdots \&Pa_{n-1}) > conf(\neg Pa_n, Pa_1\& \cdots \&Pa_{n-1})$, but the difference decreases with increasing λ and vanishes in the limit. So, also within a restricted framework of logical probabilities the justification of induction remains underdetermined.

This holds the more for standard approaches to confirmation. These approaches are much more relaxed regarding constraints of $conf$. The most fundamental constraint they put forward is that $conf$ is based on a probability function Pr satisfying the axioms of probability theory. One famous confirmation measure, namely absolute confirmation, simply identifies $conf(H,E)$ with the conditional probability $Pr(H|E)$: $conf^{abs}(H,E)= Pr(H|E)$. Another one, namely incremental confirmation, simply considers the linear probabilistic increase or decrease of H by E: $conf^{incr}(H,E)= Pr(H|E) - Pr(H)$. There are characterisation results regarding both of them which provide more or less justification for each measure (e.g., absolute confirmation avoids the so-called *paradox of irrelevant conjunctions*, and incremental confirmation avoids the so-called *ravens paradox*—(for details see Sprenger 2016)). Besides these measures for confirmation, there is a plurality of further measures. Now, regarding the probabilistic assumption that confirmation is based on a *probability* function, one can provide several arguments for justifying the axioms of probability theory. Perhaps the most famous one is the so-called *Dutch book argument*: If one interprets *Pr* as betting odds, and if one puts forward as constraint that one should not be prone to a Dutch book, i.e. a set of fair bets whose net gains are negative, then adherence to the axioms of probability theory is necessary and sufficient for being not prone to a Dutch book (see Talbott 2008; Hájek 2005). Clearly, there is some discussion about the ends. E.g., if one stipulates as an end not only non-negative gain, but positive gain, then a further condition for *Pr* is necessary, namely regularity of *Pr*, i.e. the condition that only logical truths receive a probability *Pr* of 1—for the notion of *regularity* see chapter 11; for the constraint see (Howson 2000, p.134). However, the axioms of probability theory are so widely accepted that, at least so it seems, *de facto* people fit the ends towards the means and not vice versa as they aim at a *post hoc* rational explanation of our choice of these principles. In general, such conditions and principles are considered to be "laws of consistency" or coherence (see Howson 2000, p.134). In this sense the standard approaches to confirmation might be assigned to coherentism. With respect to the problem of induction they also remain similarly underdetermined: If we take, e.g., the absolute measure of confirmation, then, probabilistically speaking, one is completely free in choosing a probability function *Pr* such that $Pr(Pa_n|Pa_1\& \cdots \& Pa_{n-1}) < Pr(\neg Pa_n|Pa_1\& \cdots \& Pa_{n-1})$. If we consider an even stronger form of induction, namely inductive generalisation of the form $Pa_1\& \cdots \& Pa_n \mathrel{|\!\!\sim} \forall x Px$ (for high enough n), then we get by Bayes' theorem:

$$Pr(\underbrace{\forall x Px}_{H} \,|\, \underbrace{Pa_1\& \cdots \& Pa_n}_{E}) = \underbrace{Pr(E|H)}_{=1} \cdot \frac{Pr(\forall x Px)}{Pr(Pa_1\& \cdots \& Pa_n)}$$

The unconditional probabilities $Pr(\forall x Px)$ and $Pr(Pa_1\& \cdots \& Pa_n)$ are also called *prior probabilities*, since they are prior to receiving any relevant evidence (in contrast, conditional probabilities are also called *posterior probabilities*, since they are conditional some evidence and according to ordinary updating one identifies them with absolute probabilities after receiving the evidence). Now, if we consider as alternative hypothesis H' the proposition $\neg\forall x Px$, then it is easy to see that there are prior probabilities which allow for higher absolute confirmation of H' than H by E: Take, e.g., $Pr(\forall x Px) = 1/4$ and $Pr(Pa_1 \& \cdots \& Pa_n) = 3/4$, then $Pr(\forall x Px | Pa_1 \& \cdots \& Pa_n) = 1/3$, hence $Pr(\neg\forall x Px | Pa_1 \& \cdots \& Pa_n) = 2/3$, hence $conf^{abs}(H, E) < conf^{abs}(H', E)$. Similarly for an anti-inductive alternative H'': $\forall x(x \neq a_1 \& \cdots \& x \neq a_n \rightarrow \neg Px)$. Whether inductive or anti-inductive inferences are licensed depends completely on the choice of the prior probabilities. So, also in the probabilistic coherentist approach the problem of induction remains unresolved due to an underdetermination of the prior probabilities: "With no assumptions at all, we get no results. Probability theory is not magic, and in its strongest pure form, the skepticism of David Hume is unanswerable" (see Skyrms 2000, p.156). However, it is important to note that *subjective Bayesians* who do not put forward any constraints for choosing prior probabilities often see this not as a vice, but a virtue of their framework:

> "The 'synthetic' premises in a probabilistic inference are generally prior, or unconditional, probabilities, and because their exogenous nature is explicitly acknowledged within the so-called subjective Bayesian theory they are often seen as its Achilles heel. Hume's argument enables us to view them in a less unfavourable light, for it implies that some degree of indeterminacy is a natural and indeed inevitable feature in any adequate theory of valid inductive inference." (Howson and Urbach 2006, p.80)

Coherentism in Probabilistic Disguise: Old Wine in New Skins? Finally, speaking about probabilistic coherence, a short note on the dynamics of probabilism and its relation to Hume's problem is in place. In using Bayes' theorem from above we were already speaking of *prior probabilities*, meaning those probabilistic statements which are relevant *prior*, i.e. *before*, gathering evidence E. Now, probabilistic statements relevant *prior* and *posterior* gathering some evidence are usually described in form of probabilistic dynamics, speaking of a change of an epistemic agent's degrees of belief from prior Pr to posterior Pr' after receiving evidence E. How an epistemic agent should change her degrees of belief is normatively stated via so-called *rules of update*. One of the most famous rules is *Bayesian update*, stating that once one gets to know E, one's posterior Pr' should be obtained from one's prior Pr by conditionalising on E:

$$\text{For all propositions } H\text{: } Pr'(H) = Pr(H|E)$$

So, e.g., to get to know E means that $Pr'(E) = 1$—before it might have been even non-strictly disbelieved: $0 < Pr(E) < 0.5$. There are also *diachronic Dutch book arguments* which allow for justifying Bayesian conditionalisation, i.e. considering it as a coherence constraint (see Vineberg 2016, sect.4.1). Now, note that Pr are one's past degrees of belief, and Pr' are one's present degrees of belief relevant also for one's future degrees of belief. Conditionalisation demands to infer from the former the latter, and since there are coherentist arguments in favour of such an inference, one might ask whether we have a successful case of inductive reasoning from the past to the present and future? However, note also that the whole "dynamics" is only determined by *the* (*very*) prior probability distribution, i.e. one's degrees of belief prior gathering any evidence at all, and the set of evidence, so the *problem of the priors* seems to still remain and underdetermine epistemic justification. Nevertheless, if we consider the relevant update case as that of having *the* prior degree of belief Pr at $t = 1$, receiving all *the* evidence E at $t = 2$, inferring at $t \geq 2$ from Pr and E *the* posterior degree of belief Pr' by help of conditionalisation, and finally holding Pr' at $t \geq 2$, then an argument for conditionalisation seems to be also an argument for an inductive inference. The problem of the priors concerns, so to say, only the truth of the premisses of this inference, but not the validity of the inference itself (likewise as for the validity of a deductive inference it does not matter whether the premisses of the inference are in fact all true or not). So, is a dynamic Dutch book argument an argument in favour of an inductive inference? Yes, it is, but it is not without an inductive assumption: The argument shows that violating conditionalisation allows a bookie to offer bets at $t = 1$ which are licensed by Pr and buying a bet at $t = 2$ which is licensed by Pr' such that the epistemic agent is guaranteed to have a net loss at $t \geq 2$ (for details see Vineberg 2016, sect.4.1). However, that there is a guaranteed net loss at $t \geq 2$ depends on the inductive assumption that the bookie can buy back a bet at $t = 2$. So the argument validates an inductive inference by help of another one.

To summarise, we have provided approaches to the problem of induction for each positive branch of epistemic justification: foundationalism, coherentism, infinitism, and naturalised epistemology. We have seen that the discussed approaches fail to account for the problem of induction, because they either leave the notion of epistemic justification underdetermined such that neither induction nor anti-induction or both are characterised as justified or not (as *good* or *bad* inductive inferences). Or, as in the case of naturalised epistemology, they fail to account for the normative element of the notion of *justification*. We have also seen that more modern approaches as, e.g., Popper's falsificationism, Carnap's logical probabilism as well as more relaxed forms of probabilism fall short of resolving the problem. In the next section we provide a formal learning theoretical argument against the possibility of accounting for the problem of induction. This is in support of the negative branch of epistemic justification, namely scepticism. However, in the subsequent section we want to approach the problem of induction with a more positive spin by showing how a redefined epistemic end allows for engineering achievable epistemic means.

5.3 The Impossibility of Humean Justification: Absolute Learnability

Now, as we have argued in the preceding section, neither the traditional nor the modern probabilistic approaches can account for Hume's problem (and sometimes also not aim at accounting for it). Hume's problem concerns the justification of induction. Its formulation in form of two horns, namely that a deductive justification is logically impossible and an inductive justification is circular, clearly presupposes a notion of *justification* (J) which is non-trivial in the sense of being non-sceptical (EJ1) and non-circular (EJ3). However, neither the foundationalist approach (vs. (EJ2) restricted to induction) as well as the infinitist approach (vs. (EJ4) restricted to induction) sketched in the preceding section can account for it. And even if one allows for a circular coherentist notion of *justification* (vs. (EJ3) restricted to induction) then, as we have seen above, these accounts also fail. Now, one might wonder whether this failure is due to the specific approaches we were discussing or whether it is generally impossible to provide an answer to Hume's task of justifying induction, i.e. there is no fundamentalist, coherentist or infinitist approach which can account for it and so only scepticism remains (vs. (EJ1) restricted to induction). As we will see now, framing Hume's problem of induction in terms of absolute learnability in fact leads to scepticism. Note that scepticism regarding a strict justification of induction is *the* common position in epistemology and philosophy of science. So, the conclusion of our argument is in full agreement with the epistemological canon. Note further that this does not imply epistemic scepticism in general, since it is intended to show that only with regards to induction strict epistemic justification fails. And furthermore, as we show in the subsequent section, one might be an epistemic sceptic regarding one notion of *justification* (regarding one "ideal" epistemic end, absolute learnability), but an epistemic non-sceptic regarding another notion of *justification* (regarding a "realistic" epistemic end, relative learnability).

Here is how we think of Hume's problem of induction in terms of meta-induction and online learning:

Hume (1748/2007)	**Meta-Induction / Online Learning**

"When it is asked, *What is the nature of all our reasonings concerning matter of fact?* [...] it may be replied in one word, *Experience*." (p.23, 32)	Reasoning happens in online prediction games and is based on past outcomes and predictions, i.e.: $f_{l,t}$ ("conclusion") is based on $f_{i,<t}$ and $y_{<t}$ ("experience").
"It must certainly be allowed, that nature has kept us at a great distance from all her secrets, and has afforded us only the knowledge of a few superficial qualities of objects;" (pp.24f, 33f)	$\mathcal{Y}$ might be construed by 🐱.
"As to past *Experience*, it can be allowed to give *direct* and *certain* information of those precise objects only, and that precise period of time, which fell under its cognizance: But why this experience should be extended to future times, and to other objects[?]" (pp.24f, 33f)	$f_{i,<t}$ and $y_{<t}$ are given, but 🐦 is not given, i.e. it is not to be expected that (e.g. inductive) $f_{l,t}$ and y_t match.
"There are no demonstrative arguments in the case[. ...] It implies no contradiction, that the course of nature may change [... and this] can never be proved false by any demonstrative argument or abstract reasoning *à priori*." (p.25, 35)	🐱 might present to the learner $f_{i,<t}$ and $y_{<t}$ such that (e.g. an inductive) $f_{l,t}$ and y_t fall apart.

Now, to show that inductive (or any other form of) reasoning or learning is possible in the strict sense can mean two things:

> "We have said, that all arguments concerning existence are founded on the relation of cause and effect; that our knowledge of that relation is derived entirely from experience; and that all our experimental conclusions proceed upon the supposition, that the future will be conformable to the past. To endeavour, therefore, the proof of this last supposition by *probable arguments, or arguments regarding existence*, must be evidently going in a circle[.]" (see Hume 1748/2007, p.26, 35f, emphasis by us)

We understand 'probable arguments' as arguments showing that a conclusion of an inference is more probable than its negation; and we understand 'arguments regarding existence' as arguments showing that the conclusion of an inference is true. So, traditionally two ends allow for justifying an inference/learning method: the first epistemic end: reaching a true conclusion from true premisses; the second epistemic end: reaching a more probable conclusion from true premisses (in the sense that the conclusion is at least more probable than its negation). Now, only deduction allows for achieving both ends, given one's premisses are true or probable: Given some (true) experience, just by applying a deductive method one reaches a true and probable conclusion. Hence, deduction is justified according to these ends. Clearly, induction does neither allow for achieving the first, nor the second end: It may lead from true or probable premisses to false or improbable conclusions. Anti-induction (as an inductive inference in the wide sense) is in the same boat. The question is, can we discriminate better between non-deductive inferences/learning methods by stipulating different ends. A natural weakening is the following one:

(EE1) reaching a true conclusion (*the truth*) from true premisses *in the long run*

(EE2) reaching a more probable conclusion from true premisses in the sense of reaching a true conclusion more often *on average*

Note that (EE1) allows for inferring wrong conclusions from true premisses, but in the long run, i.e. at some point in the infinite series of predictions, the inference or learning method must succeed in making only true predictions, i.e. having learned the truth, the true hypothesis. (EE2) is weaker in the sense that it does not ask for predictions that are *in fact* true, but for predictions which are *on average* more often true than false. The question is, is induction a better means to achieve these ends than anti-induction? As we argue now, the answer is negative. In order to keep technicalities simple, we remain in the realm of online classification. With a great deal of technicalities these results can be also expanded to the realm of online regression.

Let us begin with (EE1): Is it possible to learn the truth in the long run? In the online learning setting G with truth $\mathcal{Y}$ and prediction or hypothesis set $\mathcal{F}$, learning the truth in the long run is equivalent with absolute online learnability, i.e. having a success rate of 1 in the long run (see section 3.3). Now, according to theorem 3.23, in the case of online classification *absolute learnability* is characterised by *realisability* of G and *finiteness* of Littlestone's dimension of $\mathcal{F}$ ($Ldim(\mathcal{F})$). So, the question is, can one share the sceptics concern and still allow for realisability and finiteness? Now, the strongest sceptic concern is that truth $\mathcal{Y}$ is simply defined as inversion of the learners prediction f_l, i.e. the strategy 🐱 simply designs a prediction game G

such that $y_t = 1 - f_{l,t}$ in the binary case or more generally $y_t \neq f_{l,t}$ in any online classification case with a 0-1 loss; with a different loss function 😈 simply states y_t such that it maximises the distance from $f_{l,t}$ according to the loss function. Since there is no qualitative difference in our results regarding binary and k-ary classification games, we stick to the binary case for the remainder of our discussion (in section 3.3 we have outlined how the binary case can be generalised to any k-ary case). A non-sceptic can only exclude this *super* sceptic case by putting forward the constraint of realisability. There is a technical possibility for a learner to implement realisability into any prediction game: For any unrealisable prediction game G with the truth $\mathcal{Y}$ and the prediction or hypothesis set $\mathcal{F}$ a learner can always consider a realisable prediction game G' with $\mathcal{Y}$ and $\mathcal{F}'$, where $\mathcal{F}'$ is the set of all possible series of event outcomes. In this way, a learner can implement realisability, but this is of course at the cost of an infinite hypothesis set $\mathcal{F}'$, and the question is whether one can still learn the truth from such an infinitely large hypothesis set.

That $\mathcal{F}'$ is infinite does not automatically imply that also $Ldim(\mathcal{F}')$ is infinite. Recall our example from section 3.3 where $\mathcal{F}$ was infinite $(f_1, f_2, \ldots)$, but each forecaster or hypothesis f_i predicted 1 exactly once namely: $f_{i,t} = 1$ iff $t = i$. For this case $Ldim(\mathcal{F}) = 1$. So, if this case is also realisable, then a learner (e.g. the standard optimal algorithm) could in principle learn the truth with one single mistake, although $\mathcal{F}$ is infinite. The question is, whether a learner who tries to implement realisability can do so by also keeping $Ldim(\mathcal{F}')$ finite, although $\mathcal{F}'$ is infinite. And for quite obvious reasons the answer is: *No*: Recall from definition 3.17 that $Ldim(\mathcal{F}')$ is the maximal depth of a decision tree which is shattered by $\mathcal{F}'$, i.e. which is such that at each leaf of the tree with depth $Ldim(\mathcal{F}')$ there are two hypotheses from $\mathcal{F}'$, one predicting 0 and one predicting 1. This means that there is a guarantee for an adversary 😈 to err a learner only until $Ldim(\mathcal{F}')$ rounds. Afterwards, there is no longer a guarantee that the adversary can err the learner without having erred also all the hypotheses in $\mathcal{F}'$, which is excluded by the realisability assumption, i.e. the assumption that at least one hypothesis of $\mathcal{F}'$ is never erred. Now, since the possible event outcomes at a round t are 0 and 1, by definition the set of all possible series of event outcomes $\mathcal{F}'$ contains for t two hypotheses f_1, f_2 which match the truth up to round $t-1$ and predict differently at round t: $f_{1,<t} = f_{2,<t} = y_t$, and $0 = f_{1,t} \neq f_{2,t} = 1$. Hence, for any $t \in \mathbb{N}$: $\mathcal{F}'$ shatters a decision tree with depth t. And hence there is no maximal integer u such that $\mathcal{F}'$ shatters a decision tree with depth u. So, $Ldim(\mathcal{F}')$ is infinite. So, adding all possible series outcomes to the hypothesis set and by this construct a realisable prediction game does not work.

Just to connect this discussion further to the traditional framing of the problem: The main sceptical assumption which underlies our reasoning is that an adversary can arbitrarily err a learner. In terms of the *logic of* *deceivability* outlined in chapter 3 this means that she can pick any events in the series and put them together in form of a decision tree such that the learner f_l always errs, regardless of whether f_l is inductive, anti-inductive or whatsoever inference method. This free choice of events and putting them together in any kind of series (in a way it is a possibility to mix up past, present, and future) encodes, so we think, exactly Hume's dictum that all events are independent of each other, that there is no deductive relation between the past, present, and future (see Hume 1748/2007):

- "As to past *Experience* [...] why this experience should be extended to future times, and to other objects, which for aught we know, may be only in appearance similar;" (p.24, 33)
- "These two propositions are far from being the same, *I have found that such an object has always been attended with such an effect*, and *I foresee, that other objects, which are, in appearance, similar, will be attended with similar effects*" (p.25, 34)
- "If there be any suspicion, that the course of nature may change, and that the past may be no rule for the future, all experience becomes useless, and can give rise to no inference or conclusion." (p.27, 37)
- "It is impossible, therefore, that any arguments from experience can prove this resemblance of the past to the future; since all these arguments are founded on the supposition of that resemblance." (p.27, 38)
- "Confess, that it is not reasoning which engages us to suppose the past resembling the future, and to expect similar effects from causes, which are, to appearance, similar." (p.29, 39)

So, we conclude that given Hume's condition that there is no deductive relation between past, present, and future, *inductively* learning the truth is impossible, even in the long run: There is no epistemic means to achieve end (EE1).

Let us come to the second epistemic end, namely reaching a true conclusion from true premisses *on average* more often than a false one (EE2). What does 'on average' mean here? It means that considering all possible series of event outcomes up to some round t, taking as input the outcomes up to round $t-1$ as premisses, leads in more than 50% of the series to a true prediction or conclusion about the outcome of round t. Now, by simple combinatorial considerations it follows that this is not possible. Again, we can make our point by considering the binary classification case. To see this, one just needs to note that of all possible series of event outcomes up to round t, 50% are indistinguishable regarding the event outcomes up to round $t-1$ and differ with regards to round t in stating an outcome 0 or

1. For this reason, whichever way an inference rule f_l ends up with her prediction about event Y_t based on $y_1, \dots, y_{t-1}$, in 50% of all possible series of event outcomes up to round t it will be wrong, i.e. $f_{l,t} \neq y_t$. Note also that f_l will be right in 50% of the cases, so, averaging over all possible series of event outcomes a learning method f_l does not perform better or worse than a random choice. Table 5.3 illustrates this fact.

t	1	2	
1	0	0	50%
2	1	0	
3	0	1	50%
4	1	1	

t	1	2	3	
1	0	0	0	50%
2	1	0	0	
3	0	1	0	
4	1	1	0	
5	0	0	1	50%
6	1	0	1	
7	0	1	1	
8	1	1	1	

t	1	2	3	4	
1	0	0	0	0	50%
2	1	0	0	0	
3	0	1	0	0	
4	1	1	0	0	
5	0	0	1	0	
6	1	0	1	0	
7	0	1	1	0	
8	1	1	1	0	
9	0	0	0	1	50%
10	1	0	0	1	
11	0	1	0	1	
12	1	1	0	1	
13	0	0	1	1	
14	1	0	1	1	
15	0	1	1	1	
16	1	1	1	1	

⋱

Table 5.3: Illustration of the impossibility of achieving (EE2) in the binary classification case; left: considering 2 rounds, there are 4 prediction methods possible which take the true outcome of round 1 as premiss and have a prediction of the outcome at round 2 as conclusion; each of them fails in 50% of the cases; considering 3,4,... rounds preserves this property of failing in 50% of the cases.

So, again we have to conclude that also for the (in the long run) weaker epistemic end (EE2) there is no epistemic means to achieve it. Much more general versions of this kind of impossibility result have been also called *no free lunch results or theorems*, because these kinds of results state that without any restriction of the set of possible series of event outcomes no learning algorithm is better off than any other (the expression 'no free lunch' is due to David Haussler (see Wolpert 1996, p.1343)). Since aiming at good performance in averaging across all possible event series is sometimes also called *bias-free learning*, these results are sometimes also described as an impossibility of bias-free learning:

> "The Futility of Bias-Free Learning[:] The above discussion

> illustrates a fundamental property of inductive inference: *a learner that makes no a priori assumptions regarding the identity of the target concept has no rational basis for classifying any unseen instances.*" (see Mitchell 1997, p.42)

Such no free lunch *theorems* have been proven also for online regression and more general cases (see Wolpert 1996; Wolpert and Macready 1997). And already David H. Wolpert, who was the first to proof such a theorem, linked his result to Hume's problem by citing the central passage of the *Treatise* (see Wolpert 1996, p.1341). Interpreting the no free lunch theorems as formal explication of Hume's problem is on its way to become philosophical and machine learning folklore (see the easy to follow and nice linking of Hume's problem of induction to no free lunch theorems in Domingos 2015, chpt.3); also Schurz (2017, sect.9.1) prominently discusses no free lunch theorems in the context of Hume's problem.

Now, one way to resolve the problem is the restate (EE2) such that truth is not the aim on average across all possibilities, but across a restricted subset thereof. For practical considerations it might be reasonable to consider not all possibilities but, e.g., only those that resemble past outcomes. However, from an epistemic stance this amounts to assuming already that some principle of induction is justified. For this reason we are taking another line in the next section by restating the learning target: The epistemic end we aim at is not to achieve absolute learnability and also not absolute learnability on average, but to achieve relative learnability.

5.4 The Meta-Inductive Approach: Relative Learnability

In section 5.2 we have seen that traditional as well as common modern approaches to the problem of induction fail. By help of the learning theoretical arguments presented in section 5.3 the commonly acknowledged claim that this failure is not inherently due to these approaches, but that Hume's problem of induction cannot be resolved in principle, was framed in the setting of online learning: Putting forward the epistemic end for inferences to achieve the truth in the long run or on average is an end impossible to be achieved by any epistemic means (other than deduction). However, in the spirit of epistemic engineering, impossibilities are no dead ends, but points of departure. This is exactly what Reichenbach's so-called *pragmatic justification* or *vindication* of induction is about (see W. Salmon 1963). For Reichenbach, Hume's observation was boon and bane at the same time:

> "It seems to me that after his brilliant criticism of induction, the merits of which can not be overestimated, Hume ran the problem into a side track by his defense of inductive belief as a habit.

> [...] I do not think [one] would ever have written a paper on the habit of the syllogism. Although syllogistic inference is a habit also, as well as inductive inference, nobody would mention this fact within a logical analysis. Unfortunately, ever since David Hume's turning of the problem of the inference into the problem of a habit logicians have shared his escape from logic into psychology." (see Reichenbach 1940, p.99)

Instead of turning to psychologism or naturalised epistemology, Reichenbach prominently suggested to stipulate a new epistemic end for justifying induction:

> "Hume demanded too much when he wanted for a justification of the inductive inference a proof that its conclusion is true. What his objections demonstrate is only that such a proof cannot be given. We do not perform, however, an inductive inference with the pretension of obtaining a true statement. What we obtain is a wager; and it is the best wager we can lay because it corresponds to a procedure the applicability of which is the necessary condition of the possibility of predictions. To fulfill the conditions sufficient for the attainment of true predictions does not lie in our power; let us be glad that we are able to fulfill at least the conditions necessary for the realization of this intrinsic aim of science." (see Reichenbach 1938, pp.365f)

To make a prediction by help of a method which is truth preserving, which achieves the truth in the long run, or perhaps also which achieves the truth at least on average is sufficient for epistemic, i.e. predictive, success. However, is it also necessary? Herbert Feigl, who argued also in line with Reichenbach, stated that our ordinary notion of *justification* allows also for other ways of justifying:

> "The word 'justification' shares some of the ambiguities of the word 'reason' (as used in phrases like 'giving reasons'). As we proceed we shall find it not only indispensable but also highly clarifying to distinguish between justification in the sense of *validation* and justification in the more usual sense of an argument concerning *means* with respect to *ends*. The type of justification which we wish to distinguish from validation may be called 'pragmatic' or 'instrumental' justification (*justificatio actionis* as contrasted with *justificatio cognitionis*). [...] We shall take the terminological liberty of using the term 'vindication' as a short expression for this second meaning." (originally in Feigl 1950; reprinted as Feigl 1981, pp.239f)

So, the idea is to justify induction by showing that it is necessary for predictive success, rather it being sufficient: If any non-deductive method is predictive successful, then also induction is successful (this leaves it open whether there are any successful non-deductive methods at all). Note that this shift in the epistemic end represents a shift from absolute to relative learnability: We no longer aim at absolutely learning the truth, i.e. learning it in the long run or at least on average. Rather, we aim at learning the truth at least as good as any other method might learn it, i.e. we aim at relative learnability. Skyrms (2000, p.46) puts the idea as follows: "Our decisions are a gamble and if no method is guaranteed to be successful, then it would seem rational to bet on that method which will be successful, if any method will." He also provides a nice example illustrating the rationale behind pragmatic justification (here is a slight modification): Suppose there is a box with red, yellow, and green lights, and you are to bet on one of the colours by your life. You know there are five states possible: no lights are on, all lights are on, only the red light is on, the red and yellow lights are on, or the red and green lights are on. Note, you have to bet on one of the colours, so, if no light turns on, then you loose your life anyway. However, whenever you are successful in your prediction (and saving your life), you would have been also successful by predicting red. Hence, predicting red is not sufficient for success, but necessary in the sense that whenever you are successful with your prediction, you would have been also with predicting red.

Here is Reichenbach's argument for why induction in form of the so-called *straight rule* (see the definition in section 5.2, p.158) is necessary for epistemic success: In his very influential "*Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge*" (1938) he argues as follows:

1. "If we cannot realize the sufficient conditions of success, we shall at least realize the necessary conditions. If we were able to show that the inductive inference is a necessary condition of success, it would be justified; such a proof would satisfy any demands which may be raised about the justification of induction." (p.348)

2. "Let us introduce the term "predictable" for a world which is sufficiently ordered to enable us to construct a series with a limit." (p.350)

3. "The principle of induction [i.e. the straight rule which transfers the observed frequency to the limit] has the quality of leading to the limit, if there is a limit [i.e. if the world is predictable]." (p.353)

4. "But is it the only principle with such a property? There might be other methods which also would indicate to us the value of the limit. [...] Imagine a clairvoyant who is able to foretell the value p of the

limit in such an early stage of the series [where the straight rule still fails];" (p.353)

5. "The indications of the clairvoyant can differ, if they are true, only in the beginning of the series, from those given by the inductive principle. In the end there must be an asymptotical convergence between the indications of the clairvoyant and those of the inductive principle." (p.354)

6. "If there is any method which leads to the limit of the frequency, the inductive principle will do the same;" (p.355)

7. [Hence, asymptotical convergence or long run equality with the inductive principle is a necessary condition for success in predictable worlds.]

8. [Hence, the inductive principle is justified regarding predictable worlds.]

Now, in principle Reichenbach's solution to the problem of induction is quite simple, but it seems to be also narrow: If the world is predictable in the sense that for any distribution under investigation there is a limiting frequency, then a method that is defined as approaching this frequency in the limit (as, e.g., is guaranteed by the straight rule (see Howson 2000, p.72; see also "direct inference" in Sprenger 2016, sect.3)), will "lead to the limit". It is clear that the whole analytical argument is based on the specific interpretation of 'a series is predictable' as 'there exists a limit of the series' (see premise 2) and that by this the meaning of 'induction' is some kind of "smuggled into" the meaning of 'prediction'.

However, one might be predictive successful regarding a series of events, although the series is not predictable in the sense of having a limit. Reichenbach considered this case and concedes that one might object "by the construction of a world in which there is no series having a limit. In such a world, so our adversary might argue, there might be a clairvoyant who knows every event of a series individually, who could foretell precisely what would happen from event to event—is not this "foreseeing the future" without having a limit of a frequency at one's disposal?" (Reichenbach 1938, p.358). So, if induction succeeds only in predictable worlds, but a clairvoyant is epistemically successful also in non-predictable worlds, how can induction be necessary for epistemic success? Reichenbach's response to this challenge is ingenious and simple at the same time: He argues that whenever there is a method which is successful in making predictions, then at least the method's success is predictable, which means that one can make "an inductive inference as to the reliability of the prophet, based on his successes" (Reichenbach 1938, p.359). So, the idea is that either (i) the series under investigation is predictable in the sense of having a limit or (ii) it is not predictable. If it is, then induction in form of the straight rule predicts this limit and hence succeeds. If it is not, then either (ii.i) there is no successful prediction method at all or (ii.ii) there is at least one such prediction method. In case (ii.i) induction fails, similarly as one fails in predicting a colour of the box when no light at all is on. In case (ii.ii) the series of success of the prediction method has a limit, hence induction in form of the straight rule can be applied there (see the interpretation of Reichenbach in Skyrms 2000, p.47). Hence, induction succeeds also in case (ii.ii). So, whenever a prediction method succeeds (i.e. cases (i) and (ii.ii)), then also the inductive method succeeds. This seems to vindicate induction.

However, there is the following problem with this justification (see Skyrms 2000, pp.47f): Success in case (i) differs from that in case (ii.ii). In case (i) the inductive method is successful in predicting the limit of the *event series*. In case (ii.ii) the inductive method is shown to be successful only in predicting the limit of the *success series* of another prediction method. Furthermore, the other prediction method is successful in predicting the *event series*. Hence, in order to argue for the inductive method as being necessary for success—in the sense that if any method is successful regarding predictions of the *event series*, then also the inductive method would be successful regarding predictions of the *event series*—, one needs to show that the inductive method will be also successful in predicting the *event series*. So, what Reichenbach's argument shows for case (ii.ii) is that if any method is successful in predicting *events*, then induction is successful in predicting *success*. What is missing is to show that also in case (ii.ii) it holds: If any method is successful in predicting *events*, then induction is also successful in predicting *events*. As it turns out, the learning theory outlined in chapter 3 allows for adding this missing link between being successful in predicting *success* to being relatively successful in predicting *events*.

Exactly this programme was taken up and carried out by Schurz. In Schurz (2008b) the problem of induction is tackled based on own results and results which were ingeniously transferred from online learning theory to the philosophical debate. The idea is as follows: In the line of Reichenbach's argument, an inductive method is applied to the success rates of prediction methods (meta-induction). Then, exceeding Reichenbach's argument, a prediction about the events is made by taking into account the prediction methods' forecasts based on inductively inferred predictive success. Finally, it is shown that such a meta-prediction method allows for relative learnability. Hence, if any prediction method is successful in predicting *events*, then also this meta-inductive prediction method will be successful in predicting *events*. So, this form of meta-induction is necessary for epistemic success regarding the prediction of *events* in the sense that if any prediction method is successful in predicting *events*, then also the meta-inductive method will be.

Here are the details: Let G be a regression game with the truth $\mathcal{Y}$ and the prediction methods $\mathcal{F}$ (for classification games the same argument holds with respect to different forms of success as outlined in chapter 4). Then we can define, in the line of argumentation suggested by Reichenbach, an inductive learning method on the series of success rates of $f_i \in \mathcal{F}$ simply by applying the straight rule to this series. For each f_i we define a success learner f_{ls_i} as follows:

$$f_{ls_i,t} = succ_{i,t-1}$$

Intended is the interpretation of $f_{ls_i,t}$ as the relative success learner's prediction of the relative success rate of f_i at t, and as the definition shows, this prediction simply consists in transferring the past relative success rate ($t-1$) to the future (t). Clearly, if f_i is predictively successful to some degree r, i.e. if $\lim_{t\to\infty} succ_{i,t} = r$, then the success learner f_{ls_i} will learn and predict this: $\lim_{t\to\infty} f_{ls_i,t} = r$. Hence, f_{ls_i} will be successful in predicting *success*. So much for Reichenbach's argument.

Now, we go on with the extension of Schurz (2008b, sect.7), showing that based on this inductive inference we can define a method which allows also for being successful in predicting the *events* of G. We can do so by defining the attractivity or relative success based weighting meta-inductive method f_{ami} as defined in definition 3.39. Recall that $f_{ami,t}$ is a weighted average of the prediction on Y_t of $f_i \in \mathcal{F}$, where the weights $w_{i,t}$ are normalised attractivities, i.e. relative successes (more specifically: attractivities):

$$f_{ami,t} = \sum_{f_i \in \mathcal{F}} w_{i,t} \cdot f_{i,t}$$

Now, the attractivity based forecasting method f_{ami} is a meta-method, inasmuch as it is defined on the basis of the predictions of other methods. It is an inductive method, inasmuch as it infers the weights for the future prediction $f_{ami,t}$ from past relative success:

$$\begin{aligned} w_{i,1} &= \frac{1}{n} \\ w_{i,t} &= \begin{cases} \frac{max(0, f_{ls_i,t} - succ_{ami,t-1})}{\sum\limits_{j=1}^{n} max(0, f_{ls_j,t} - succ_{ami,t-1})} & \text{if the denominator} > 0 \\ \frac{1}{n} & \text{otherwise} \end{cases} \end{aligned}$$

Given our interpretation of the success learner f_{ls_i} from above we can now interpret the weights as f_{ami}'s prediction about the normalised attractivities, i.e. relative success rates, of the prediction methods in $\mathcal{F}$. So we can say that at each round where no information about the attractivities is available as, e.g., in round 1, f_{ami} applies a principle of indifference in her prediction of the value $w_{i,t}$. And at all the other rounds f_{ami} applies indirectly the straight rule via setting $f_{ls_i,t} = succ_{i,t-1}$. Figures 5.3 and 5.4 provide an example for f_{ami}'s success dependent choice of such a prediction.

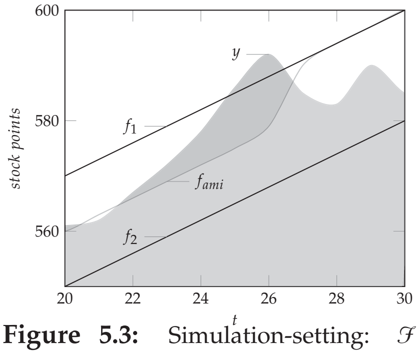

Figure 5.3: Simulation-setting: $\mathcal{F} = \{f_1, f_2\}$; since f_1 and f_2 are at the beginning equally near to the truth, f_{ami} weights them equally. From day 23 on f_1's prediction is more accurate than that of f_2. Nevertheless it takes f_{ami} three more days until it weights f_1's prediction higher than that of f_2, because until this time both competitors had lower success rates than f_{ami}.

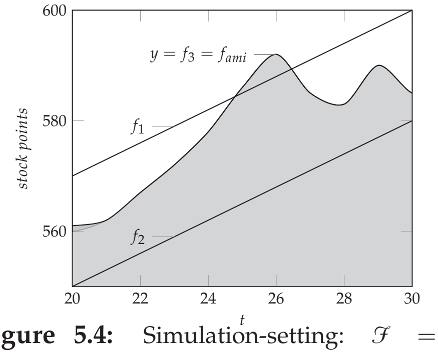

Figure 5.4: Simulation-setting: $\mathcal{F} = \{f_1, f_2, f_3\}$; at day 20 f_{ami} starts with the average of its competitor's prediction. Since from the beginning on only f_3's success rate is equal to or better than that of the meta-inductive method, f_{ami} sticks also at the following days to the correct prediction.

(In both figures y are the stock points of AAPL *Apple Inc., NasdaqGS* (November 2012). f_1 and f_2 are feigned trend lines of the stock, using only preceding chart information of `http://www.nasdaq.com/symbol/aapl/`. Simulation was performed with scripts of the language *PERL*.

Now, from theorem 3.40 we know that f_{ami} is a no regret learning algorithm, since the regret is $aregret_{\langle ami,i\rangle,t} \leq \sqrt{n \cdot t}$ for any $f_i \in \mathcal{F}$, and hence the per round regret $\leq \sqrt{n \cdot t})/t$ vanishes in the limit. So, whenever there is an f_i which is predictively successful to some degree r, i.e. if $\lim_{t\to\infty} succ_{i,t} = r$, then also f_{ami} will be successful at least to degree r:

$$\lim_{t\to\infty} succ_{ami,t} \geq r$$

This adds the missing link to Reichenbach's vindication of induction: Whenever there is a method which is in the long run to some degree successful in predicting *events*, then meta-induction is also in the long run at least to the same degree successful in predicting these *events*. So, meta-induction allows for relative learnability of $\mathcal{F}$. In this sense, meta-induction is also necessary for predictive success regarding *events*.

So, we see that if we put forward the epistemic aim of relative learnability, then meta-induction is an epistemic means to achieve this end. But what about object-induction? Can one also show that object-induction in the sense of enumerative induction or the straight rule applied to series of events directly (instead of success rates) is justified? Yes, one can, but in a much weaker sense. The "Reichenbach-Schurz" approach to Hume's problem of object-induction is twofold. Roughly it is as follows (see Schurz 2008b, p.282): First, it is argued by deductive means as above that meta-induction allows for relative learnability of $\mathcal{F}$. Hence, predicting in accordance with meta-induction will be long run optimal. Second, taken for granted the past success of classical inductive methods—something that is, e.g., also not scrutinised by Hume himself (see Howson 2000, p.4)—it follows that also selecting these methods for predictions of unobserved data and events allows for optimal predictive success. This justification of object-induction is weaker than that of meta-induction, inasmuch meta-induction is an epistemic means that definitely leads to the epistemic end of relative learnability. However, object-induction is such a means only if its success rate in fact outperforms that of all alternatives, and this is a fact which might change (for this reason Schurz 2008b, p.304, speaks of an 'a posteriori' justification of induction). So, the meta-inductive solution to the problem of epistemically justify (J) induction is not an absolute, but a time-dependent one; sloppily: Up to now object-inductive methods proved to be successful and hence by meta-induction we are justified in applying them; however, caution, this might change at some point in time!

To sum up this chapter, we saw that Hume's problem of induction in the sense of proving absolute learnability by help of induction cannot be resolved for principal reasons. From an epistemic engineer's stance one therefore has to redefine the epistemic goal in question. We have seen that Reichenbach provided a reasonable redefinition in form of a *vindication* of induction meaning that inductive methods allow for relative learnability in the sense of being optimal in the long run. However, in Reichenbach's original approach an important link was missing, namely a proof that successful induction over success rates can be also cashed out for successful induction over events. We have seen that Schurz' approach allows for adding this missing link and that by this induction can be justified in the sense of allowing for relative learnability.

If we were concerned with Hume only, this would be a nice point to stop. However, in the twentieth century a new problem of induction was put forward, namely Nelson Goodman's so-called *new riddle of induction*, which puts forward a problem for the justification of induction from a new angle. In the next chapter we investigate this problem, its impact to the "Reichenbach-Schurz" approach to induction and how it might be overcome.

Chapter 6

Induction and Goodman's New Riddle

This chapter discusses Goodman's new riddle of induction *and shows that it is a problem also for meta-induction. Afterwards, underlying assumptions are curved out and further investigated. Finally, it is argued that these assumptions are incoherent for which reason the new riddle of induction looses its anti-inductive force.*

In the preceding chapter we have argued that once one modifies the task of epistemically justifying induction from proving its ability of absolute learnability to proving its ability of relative learnability, one can account for the problem of induction: By help of meta-induction over success rates of object methods one can learn the best methods and achieve long run optimality in general. We argued that in a second step one can justify object-induction by using meta-induction and selecting object-induction in accordance with its past success. In particular, we can justify enumerative induction by help of its past success.

However, as Goodman has shown already as early as 1946, Hume's problem of induction is accompanied by a new problem, the so-called *new riddle of induction*. According to the new riddle, one might be able to resolve somehow Hume's problem of induction by, e.g., restating the epistemic end of justification and in this way provide a justification for enumerative induction. However, by *equivalent rephrasing* the induction basis one can use the justified inference method in order to also justify anti-induction (see Goodman 1946, 1955/1983, chpt.3). So, Goodman's new riddle tightens Hume's problem: Whichever solution you come up with in order to justify induction, the same solution allows also for justifying anti-induction. As long as this problem remains unresolved, any proposal to the problem of induction allows for justifying contradicting conclusions.

In this chapter we tackle Goodman's new riddle. First, we present the new riddle in section 6.1. Afterwards, in section 6.2, we aggravate the problem for the approach to induction by help of meta-induction in showing that the same problem shows up on the meta-level too and seems to allow for justifying meta-anti-induction. In section 6.3 we carve out the underlying assumptions and general constraints of this problem, namely the problem of language dependency and constraints of language independency. Finally, in section 6.4 we provide a new solution to the problem of language dependency and show how this solution provides an answer also to Goodman's new riddle.

6.1 The New Riddle of Induction

Let us briefly recap the meta-inductive justification of induction and why meta-induction seems to rule out such a justification for anti-induction. For illustrative purposes we will concentrate on enumerative induction and anti-induction. In section 5.2 we described enumerative induction $\mid\!\sim$ via the schema:

$$Pa_1, \ldots, Pa_{n-1} \mid\!\sim Pa_n$$

And enumerative anti-induction $\mid\mid\!\sim$ via the schema:

$$Pa_1, \ldots, Pa_{n-1} \mid\mid\!\sim \neg Pa_n$$

If we consider the problem of justifying induction in comparison to anti-induction, we can do so by help of a prediction game G with basically two prediction methods in $\mathcal{F}$: f_1 which predicts in accordance with $\mid\!\sim$ and f_2 which predicts in accordance with $\mid\mid\!\sim$. Now, assume that the event Y of interest is about the colour of emeralds (let us assume that an emerald is not by definition green), and that the colour space to be predicted is binary ($\{0, 1\}$): green (P) vs. blue ($\neg P$). The observational basis, which is the available past experience or event outcomes, grows with each round and let us assume it is as follows: From round 2 with $\{Pa_1\}$ to round 3 with $\{Pa_1, Pa_2\}$ to … to round n with $\{Pa_1, \ldots, Pa_{n-1}\}$. Now, given this series, clearly at each round up to n induction f_1 succeeded and by this gained full success, whereas anti-induction f_2 failed at each round up to n and by this gained zero success. The past success of f_1 over that of f_2 makes f_1 fully attractive to the meta-inductive learner f_{ami}. Since f_{ami} is guaranteed to be long run optimal (i.e. allows for relative learnability), and f_{ami} fully weights f_1's predictions, f_1 is justified given its past success. f_2 lacks such a justification, because f_{ami} ignores f_2's predictions. So, according to meta-induction's optimality (f_{ami}) and object-induction's past success (f_1) as well as object-anti-induction's past failure (f_2), induction is justified over anti-induction.

Now, Goodman (1946) presented a method which allows for turning such a justification of induction to a justification of anti-induction. We can illustrate the method as follows (see Goodman 1955/1983, chpt.3; see also the discussion in Cohnitz and Rossberg 2006, chpt.2; and Thorn 2018, sect.1): We described the prediction problem with the basic expressions 'green' (P) and 'blue' ($\neg P$). However, we might also describe the prediction problem with the basic expressions 'grue' (Q) and 'bleen' ($\neg Q$), where 'grue' is defined as 'green until some round $n-1$ and blue starting with n', and 'bleen' is defined as 'blue until the same round $n-1$ and green starting with n':

$$Qx \quad \Leftrightarrow_{df} \quad (Px \leftrightarrow (x = a_1 \vee \cdots \vee x = a_{n-1}))$$

Note that given these expressions, the observational basis from before remains structurally unchanged: $\{Qa_1, \ldots, Qa_{n-1}\}$. Now, the above definition of 'grue' (Q) by help of 'green' (P) is logically equivalent with the following definition of 'green' by help of 'grue':

$$Px \quad \Leftrightarrow_{df} \quad (Qx \leftrightarrow (x = a_1 \vee \cdots \vee x = a_{n-1}))$$

So, if we translate the predictions of f_1 and f_2 into 'grue'/'bleen' statements, then this amounts to f_1 predicting $\neg Qa_n$ (the next emerald will be not grue, i.e. bleen) and f_2 predicting Qa_n (the next emerald will be grue). If we consider induction applied on 'green'/'blue' statements as justified, and if we allow for definitional transformations among justified inferences (i.e. justification is invariant under definitional transformations), then we need to consider also f_1's prediction as justified. However, note that f_1's conclusion is $\neg Qa_n$, although the observational basis is $\{Qa_1, \ldots, Qa_{n-1}\}$. Hence, f_1's prediction is an anti-inductive inference. More generally: If justification (J) is preserved among definitional transformations, then any justification of induction can be transformed also to a justification of anti-induction. Table 6.1 illustrates this schema.

$$\begin{tikzcd}
Pa_1, \arrow[d, Leftrightarrow, "df"'] & Pa_2, \arrow[d, Leftrightarrow, "df"'] & \cdots, & Pa_{n-1} \arrow[d, Leftrightarrow, "df"'] & \mid\!\sim & Pa_n \arrow[d, Leftrightarrow, "df"'] & J \arrow[d, Rightarrow] \\
Qa_1, & Qa_2, & \cdots, & Qa_{n-1} & \mid\mid\!\sim & \neg Qa_n & J
\end{tikzcd}$$

Table 6.1: Schema of Goodmanian justification (J) of anti-induction $\mid\mid\!\sim$ by help of a justification for induction $\mid\!\sim$: If inductive inferences are justified and justification is invariant under definitional transformations, then also an anti-inductive inference can be justified by switching the language (from P-expressions to Q-expressions).

Note further, that the situation seems to be even worse: By applying the same method twice one can show that any justification of induction allows for justifying contradicting statements. To see this, take the example from above and note that we have $\{Pa_1, \ldots, Pa_{n-1}\}$, hence also $\{Qa_1, \ldots, Qa_{n-1}\}$. Now, by applying induction on both observation bases we justifiably end up with Pa_n and Qa_n. By the definition above we can transform Qa_n to $\neg Pa_n$. So, under the assumption that justification is invariant under definitional transformation, we get also justification for $\neg Pa_n$. Hence, both, Pa_n as well as $\neg Pa_n$ are justified by help of induction *and* definitional transformation.

So, not only the versions of infinitism and coherentism we presented in section 5.2 were prone to provide a justification for anti-induction in the same way they provide a justification of induction. Rather, as Goodman's argument shows, any approach of justifying induction seems to be prone to this problem: Every justification of induction can be turned to a justification of anti-induction. However, note that the meta-inductive justification of object-induction is not a justification *per se*, i.e. an unconditional justification, but a justification conditional on the object-inductive method's past success. And so in order for Goodman's problem to show up for the meta-inductive justification of object-induction, object-anti-induction needs to be also shown to be predictively successful before it can be considered as conditionally justified—conditional on its past success. As one can see from the example above, this is not automatically guaranteed: If we consider the cumulative success of the methods and their definitional translated pendants, then one can see that the inductive method f_1 is successful, whereas the anti-inductive method f_2 fails to be successful. In order to express this fact exactly we would need to modify these methods in order to cover also mixed cases where the premise set contains not only either positive or negative instances, but also a mixture of both of them. However, if we consider large enough observational bases and allow for neglecting a marginal number of counterexamples, then we can illustrate this fact as schematically expressed in table 6.2: What at round n appears as a P-anti-inductive definitional transformation of P-induction is, considering large enough observational bases approximative Q-induction. And likewise: What at round n appears as a Q-inductive definitional transformation of P-anti-induction is under the same assumption approximative Q-anti-induction. Hence: In the long run induction is predictively successful (under both descriptions: P as well as Q statements) and anti-induction is predictively unsuccessful (also under both descriptions). So, we can conclude that, at least regarding the example provided by Goodman, meta-induction still justifies object-induction and not object-anti-induction, since its justification is conditional on past success and in the long run object-induction is successful whereas object-anti-induction fails.

So, does meta-induction provide not only a solution to Hume's *old* problem of induction, but also to Goodman's *new* riddle? As we will show in the next section, this is not the case: Goodman's new riddle shows up also for the meta-inductive solution to the problem of induction.

	P-induction	P-anti-induction	
$\sum s_1$			$\sum s_2$
1	$Pa_1 \mid\!\sim Pa_2$	$Pa_1 \mid\!\sim \neg Pa_2$	0
2	$Pa_1, Pa_2 \mid\!\sim Pa_3$	$Pa_1, Pa_2 \mid\!\sim \neg Pa_3$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$Pa_1, \ldots, Pa_{n-1} \mid\!\sim Pa_n$	$Pa_1, \ldots, Pa_{n-1} \mid\!\sim \neg Pa_n$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Q-induction (approximative)	Q-anti-induction (approximative)	
$\sum s_1$			$\sum s_2$
1	$Qa_1 \mid\!\sim Qa_2$	$Qa_1 \mid\!\sim \neg Qa_2$	0
2	$Qa_1, Qa_2 \mid\!\sim Qa_3$	$Qa_1, Qa_2 \mid\!\sim \neg Qa_3$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$Qa_1, \ldots, Qa_{n-1} \mid\!\sim \neg Qa_n$	$Qa_1, \ldots, Qa_{n-1} \mid\!\sim Qa_n$	0
$n+1$	$Qa_1, \ldots, Qa_{n-1}, \neg Qa_n \mid\!\sim \neg Qa_{n+1}$	$Qa_1, \ldots, Qa_{n-1}, \neg Qa_n \mid\!\sim Qa_{n+1}$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n+m$	$\ldots, \neg Qa_n, \ldots, \neg Qa_{n+m-1} \mid\!\sim \neg Qa_{n+m}$	$\ldots, \neg Qa_n, \ldots, \neg Qa_{n+m-1} \mid\!\sim Qa_{n+m}$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 6.2: Meta-inductive solution to Hume's problem with a solution to Goodman's riddle as a byproduct? Upper left part: P-induction on the basis of P statements. Lower left part: definitional transformations of these predictions. Upper right part: P-anti-induction on the basis of P statements. Lower right part: definitional transformations of these predictions. Given a fixed switching point n for the definition of 'grue' (Q) by help of 'green before n' (P) and 'blue at and after n' ($\neg P$), a definitional transformation of P-induction turns out to approximate ($m \gg n$) also Q-induction, whereas such a transformation of P-anti-induction approximates also Q-anti-induction. Note that induction is successful (cumulative success on the left side), whereas anti-induction fails to be successful (cumulative success on the right side). This is the reason why in this example the meta-inductive justification of induction cannot be used as a justification of anti-induction.

6.2 The New Riddle of Meta-Induction

Recall, Goodman's recipe for cooking up a justification for anti-induction given a justification for induction consists in finding a definitional translation which transforms an inductive inference in an anti-inductive one. We have seen in the example of the preceding section that a transformation of 'green'/'blue' statements to 'grue'/'bleen' statements served this purpose. Under the assumption that justification is preserved under such definitional transformations, also anti-induction becomes justified.

In the meta-inductive justification of induction a straightforward application of Goodman's translations failed, since for any switching point and high enough round numbers a 'green'/'blue'-inductive inference is transformed to an (approximate) 'grue'/'bleen'-inductive inference and analogously for the anti-inductive inference. For this reason the predictive success of the 'green'/'blue'-inductive inferences also centres around 'grue'/'bleen'-induction and hence also in the 'grue'/'bleen' setting induction is justified by the meta-inductive algorithm. However, there are other definitional transformations which allow for justifying anti-induction. One possibility is discussed, e.g., in (Schurz 2019, sect.9.2.5) where so-called *Goodman methods* are introduced which work with more than one switching point (in principle there can be arbitrarily many switching points). We aim at a simpler definitional transformation which allows not only for justifying object-anti-induction, but also meta-anti-induction by help of meta-induction.

The idea is as follows: In order to justify object-anti-induction we first justify meta-anti-induction by help of meta-induction, and then object-anti-induction by help of meta-anti-induction. In order to justify meta-anti-induction we are not after definitional transformations of event predictions. Rather we are after definitional transformations of success itself. It turns out that there is a very simple definitional transformation, namely that of success to loss and vice versa. We do not want to mess up our established terminology and since we aim at inverting the success by help of definitional transformation, we will speak of 'score' $\mathscr{s}$ in the non-Goodman case and of 'erocs' $\reflectbox{$\mathscr{s}$}$ in the Goodman case. Now, the definitional translations are as follows: We define 'erocs' by help of 'score' via:

$$\reflectbox{$\mathscr{s}$}_{i,t} =_{df} 1 - \mathscr{s}_{i,t}$$

And we can define 'score' by help of 'erocs' (arithmetically) equivalently as:

$$\mathscr{s}_{i,t} =_{df} 1 - \reflectbox{$\mathscr{s}$}_{i,t}$$

Hence, $\mathscr{s}$ and $\reflectbox{$\mathscr{s}$}$ are interdefinable. Note also that by definition 2.9 of the score: $\mathscr{s}_{i,t} = 1 - \ell_{i,t}$, so we get: $\reflectbox{$\mathscr{s}$}_{i,t} = \ell_{i,t}$. We also want to define the Goodmanian version of loss which we call 'ssol' $\reflectbox{$\ell$}$:

$$\reflectbox{ℓ}_{i,t} =_{df} 1 - \reflectbox{$\mathscr{s}$}_{i,t}$$

Hence, $\reflectbox{$\ell$}_{i,t} = \mathscr{s}_{i,t}$. So, also ℓ and $\reflectbox{$\ell$}$ are interdefinable. We now can justify meta-anti-induction by help of the justification of meta-induction in accordance with the schema in table 6.3.

For simplicity reasons, let us consider the exponential weighting meta-inductive prediction method f_{emi} as discussed in section 3.4. It allows for relative learnability (theorem 3.37), and hence is access optimal. Hence it is also epistemically justified. Now, according to definition 3.34, f_{emi} predicts according to a weighted average of the predictions of the object-level methods:

$$f_{emi,t} = \sum_{i=1}^{n} w_{i,t} \cdot f_{i,t}$$

$$\begin{tikzcd}
s \arrow[r, Rightarrow] \arrow[d, Leftrightarrow, "df"'] & f_{mi} \arrow[r, Rightarrow] \arrow[d, Leftrightarrow, "df"'] & \text{optimality} \arrow[r, Rightarrow] & J \arrow[d, Rightarrow] \\
\rotatebox{90}{s} \arrow[r, Rightarrow] & f_{m\text{¡}} & & J
\end{tikzcd}$$

Table 6.3: Schema of justifying meta-anti-induction $f_{m\text{¡}}$ by help of meta-induction f_{mi}: Taking ordinary scores s as input, meta-inductive weighting is optimal and hence justified. Now, ordinary scores can be definitionally transformed to Goodmanian versions of loss (ssol), $\rotatebox{90}{$s$}$. The Goodmanian version of meta-induction f_{mi}, namely meta-anti-induction $f_{m\text{¡}}$ uses ssol as input. Since meta-induction and meta-anti-induction are definitionally intertranslatable, and since the former is justified, also the latter seems to be justified.

The weight for an object-level predictor f_i at round t results from normalising the exponential of the negative cumulative loss with some learning parameter η: $w_{i,t} = N(\exp(-\eta \cdot \sum_{u=1}^{t-1} \ell_{i,u}))$. Since $s_{i,u} = 1 - \ell_{i,u}$, we get $w_{i,t} = N(\exp(-\eta \cdot (t - \sum_{u=1}^{t-1} s_{i,u}))) = N(\exp(-\eta \cdot t) \cdot \exp(\sum_{u=1}^{t-1} s_{i,u})^{\eta})$. So, the weights of the exponential weighting meta-inductive prediction method f_{emi} are proportional to $N(\exp(\sum_{u=1}^{t-1} s_{i,u})^{\eta})$:

$$w_{i,t} \quad \propto \quad N\left(\exp\left(\sum_{u=1}^{t-1} s_{i,u}\right)^{\eta}\right)$$

N is a normalising operation and η is a learning parameter we can get rid of by a *doubling trick*—for details see section 3.4. One can see that f_{emi} is a meta-inductive prediction method in the sense that its prediction is based on induction over cumulative success or scores: The weight for the prediction at round t depends on the cumulative success or score up to round $t-1$.

Now, we can define also a meta-anti-inductive method which bases her prediction not on induction over cumulative success or scores, but anti-induction over such success or scores. So, it weights the forecasters which scored most in past lowest and those which scored least in past highest. Note that this is equivalent to weighting the forecasters with highest cumulative loss in the past highest, and forecasters with lowest cumulative loss lowest. E.g., the anti-meta-inductive exponentially weighting prediction method $f_{em\text{¡}}$ can be defined analogously to f_{emi} simply via replacing ℓ by s in definition 3.34. Its weights are then proportional to the cumulative loss in the exponent:

$$w_{i,t} \quad \propto \quad N\left(\exp\left(\sum_{u=1}^{t-1} \ell_{i,u}\right)^{\eta}\right)$$

At this point we have all ingredients for cooking up a justification for meta-anti-induction:

1. Meta-induction f_{emi} is justified (J) by its optimality (theorem 3.37).

2. Recall, by definition of the Goodmanian version of loss, ssol $\rotatebox{180}{$\ell$}$, we know that $s_{i,t} = \rotatebox{180}{ℓ}_{i,t}$.

3. Hence, the Goodmanian definitional translation of f_{emi} is a meta-anti-inductivist $f_{em\rotatebox{180}{$i$}}$.

4. Justification J is preserved among definitional transformations.

5. Hence, also meta-anti-induction $f_{em\rotatebox{180}{$i$}}$ is justified.

By help of the conclusion of this argument (5) we can justify also object-anti-induction without any need of finding suitable definitional transformations for object-level predictions in the setting of prediction games. We can simply argue analogously to the "Reichenbach-Schurz" approach to Hume's problem of object-induction by help of meta-induction in a twofold way: First, we showed already that meta-anti-induction is justified. Second, taken for granted the past failure of object-anti-inductive methods, the meta-anti-inductive choice of object-anti-induction transmits its justification also to the object-anti-inductive method. Again, this justification of object-anti-induction is weaker than that of meta-anti-induction, inasmuch as meta-anti-induction is justified unconditional its current performance, whereas object-anti-induction is justified only as long as its success rate is in fact outperformed by all of its alternatives. Again, this fact might change.

So, we have seen that Goodman's recipe allows also for justifying meta-anti-induction as well as object-anti-induction. This seems to be a quite dissatisfying result for meta-induction as well as object-induction. And indeed, there seems to be little which can be done about it: Considering the argument from above one can see that premise 1 is the main result of section 5.4 which depends on deductive proofs, premisses 2 and 3 are about definitional translations as introduced in this section, and hence there is not much space left to escape the *justificatory inflationism*. However, premise 4 is a general constraint on justification J which needs to be further explored. As we will show in the subsequent section, this constraint can be generalised and has even more devastating consequences. However, in the final section of this chapter we will provide an argument which is intended to show that this constraint is incoherent. Hence, justification of meta-anti-induction Goodman style fails due to the falsity or inadequacy of premise 4: Justification needs not be invariant under definitional transformations. Note that Schurz (2019, sect.4.2) argues also against premise 4 by stressing the fact that in order to distinguish between induction and anti-induction we need to presuppose fixed qualitative basic expressions.

6.3 The New Riddle Generalised: Language Dependency

It is well-known in philosophy of science that some very important notions like the notion of justification (J) confirmation ($conf$), verisimilitude, simplicity etc. are often explicated in a way that makes the application of the notions to theories dependent on the language the theories are formulated in. This phenomenon was recently also observed in many other areas of philosophy and the sciences and so quite a lot of literature on this topic arose (for an introductory overview see, e.g., Miller 2006, chpt.11; and Schurz 2013, sect.5.11.3; see also Niiniluoto 1998, sect.6; Oddie 2013; Schurz and Weingartner 2010, sect.6).

In this section we intend to generalise the problem of justifying anti-induction by help of definitional transformations. Such transformations show, that the aforementioned notions depend on the language one defines them in or applies them to. For this purpose we provide an exact characterisation of the problem of language dependence in the following. In a nutshell a theory is called 'language dependent', if its main notions apply to synonymous theories differently. So, e.g., if these notions are applied in theory evaluation, an alleged underdog can quite often beat her opponent just by switching the language, which sounds at least strange and in case that the notions are widely accepted also paradox.

Now, let us begin with explicating the devastating constraint underlying the justification of anti-induction, namely the constraint that justification J needs to be invariant under definitional translations or language independent (recall premise 4 of the main argument in the preceding section). Suppose you have some competence in written Bulgarian and so you might know that 'Vizh, ima zaek!' is an adequate latinised Bulgarian translation of the English sentence 'Look, there is a rabbit!'. Suppose furthermore that indeed there is a rabbit in front of you and you advice a Bulgarian to look there by exclaiming 'Vizh, ima zaek!'. Since you are supposed to be competent in written and not spoken and gesticulated Bulgarian only, you might be surprised if she shakes her head instead of nodding. But by asking or testing her reaction on other unambiguous and obvious truths or falsities you will pretty soon figure out that the Bulgarian gestures for affirmation and negation are inverse to, e.g., the English ones. So you may learn that in general she will shake her head when you will nod yours and vice versa. This means that you can figure out a rule for intertranslating between her and your gesticulating behaviour and so both are in some way or another synonymous.

As there is nothing essential in the Bulgarians' way of affirming and negating and—believe it or not—also not in the Englander's, it makes no sense to prefer one convention against the other except for pragmatic reasons as, e.g., the fruitfulness of a convention. What holds for simple gestures of natural language holds the more for theories formulated in conventional artificial languages and so one should not bother too much about different choices of the primitive vocabulary of two different theories about one and the same domain if there are some rules for intertranslating the vocabulary and theorems of the theories. So, e.g., up to conventional and pragmatic reasons it does not matter whether one chooses the Sheffer-stroke $|$ instead of the usual connectives $\neg, \&, \vee, \rightarrow, \leftrightarrow$ (logics) or whether one chooses as primitive a conditional (Pr^2) instead of an absolute probability function Pr^1 (philosophy of science), a modality for obligation $\mathscr{O}$ instead of permission $\mathscr{P}$ (ethics), or overlapping $\circ$ instead of parthood $\sqsubseteq$ (metaphysics). Needless to say, you should also not be bothered about doing amateur mineralogy on emeralds in a 'grue'/'bleen' language instead of a 'green'/'blue' language.

Bearing this in mind it is widely accepted that if one is not interested especially in language properties like the length or complexity of formulas and also not in pragmatic properties like being easier memorisable or shorter in derivation length, then one should put intertranslatable or synonymous theories on a par. In philosophy of science, e.g., this maxim is well known as the constraint for theories to be translationally invariant or language independent (for a general discussion see, e.g., Miller 2006, chpt.11).

Let us make this constraint clear by clarifying first the notion of 'intertranslatability' or 'synonymy': Technically seen the most favourable way of translating expressions and statements is to translate them by explicit definitions. What is true on the level of single expressions and statements holds also on the level of theories, and so we will say—in accordance with a common proposal (see, e.g., Bouvère 1978; and Kanger 1968)—that two theories Φ and Ψ are intertranslatable or synonymous iff they have a common definitional extension, i.e. there are definitions for all expressions of the one in terms of the other such that the expansions of the theories by the definitions are logically equivalent. More precisely:

Definition 6.1. If the descriptive vocabulary of Φ and Ψ is disjoint, then Φ and Ψ are synonymous iff there is a theory $\mathscr{X}$ and there are sets $\mathscr{D}_\Phi$, $\mathscr{D}_\Psi$ such that

1. $\mathscr{D}_\Psi$ contains exactly one definition for each descriptive symbol in Ψ in terms of descriptive symbols of Φ only, and

2. $\mathscr{D}_\Phi$ contains exactly one definition for each descriptive symbol in Φ in terms of descriptive symbols of Ψ only, and

3. $\mathscr{X} \vdash\dashv \Phi \cup \mathscr{D}_\Psi$ as well as $\mathscr{X} \vdash\dashv \Psi \cup \mathscr{D}_\Phi$
(which is to say that: $\Phi \cup \mathscr{D}_\Psi \vdash\dashv \Psi \cup \mathscr{D}_\Phi$ (and sometimes $\mathscr{D}_\Phi \vdash\dashv \mathscr{D}_\Psi$)).

There is a quite straightforward way to expand the concept of synonymous theories to the concept of synonymous logics (see Pelletier and Urquhart 2003, pp.265ff), but since our argument is only about the constraint to treat synonymous theories on a par and not synonymous logics (the latter is almost automatically satisfied by all considered theories), we will not take into account such an expansion.

Note that the condition of a disjoint descriptive vocabulary is not necessary, but convenient here. In order to compare two theories one needs, at least in some cases, to separate the vocabulary. Take our example of the English and Bulgarian affirmation and negation: Assume that in both languages it holds that affirmation and negation are exclusive. By just defining (English) affirmation with the help of (Bulgarian) negation and vice versa without distinguishing the vocabulary one would end up with a contradiction. So one has to separate English affirmation and negation from Bulgarian one. Something similar holds, e.g., for a translation of a description of set theory using $\subset$ for being a (proper or improper) subset with a description of set theory using the same sign ($\subset$) for being a proper subset. For a technically more favourable way of guaranteeing the disjointedness of the descriptive vocabulary without conditionalising the definition of *synonymy* see (Kanger 1968, pp.2f).

The second part of our clarification concerns the notion of 'language dependence': We will say that a property or relation R^n or an operation f^n respectively of a theory Φ is language dependent iff it does not hold equally of or operates differently among synonymous theories. More precisely:

Definition 6.2. An n-ary relation R^n or operation f^n of a theory Φ is language dependent iff there are $\Phi_1, \ldots, \Phi_n$ and $\Psi_1, \ldots, \Psi_n$ such that

1. Φ_1, Ψ_1 and $\ldots$ and Φ_n, Ψ_n are synonymous (whereby it is assumed that there is an overall definitional extension), and
2. $\Phi \vdash R^n(\Phi_1, \ldots, \Phi_n)$ and $\Phi \vdash \neg R^n(\Psi_1, \ldots, \Psi_n)$
or $\Phi \vdash f^n(\Phi_1, \ldots, \Phi_n) \neq f^n(\Psi_1, \ldots, \Psi_n)$

Note that the extra constraint in condition 1 of this definition, namely that there must be an overall definitional extension, is necessary in order to have it that, e.g., the classical logical consequence relation is language independent. E.g., it holds in classical logic that $\{\varphi_1 \& \psi_1\} \vdash \{\varphi_1\}$, whereas $\{\varphi_2\} \nvdash \{\psi_2\}$. If the relata of the relation or operation were to be compared separately only, then one could provide definitions for φ_2 in terms of φ_1 and ψ_1, e.g. $\varphi_2 \leftrightarrow (\varphi_1 \ \& \ \psi_1)$, and vice versa, e.g. $\varphi_1 \leftrightarrow \varphi_2$, $\psi_1 \leftrightarrow \varphi_2$ such that $\{\varphi_1 \ \& \ \psi_1\}$ is synonymous to $\{\varphi_2\}$, and one could also provide separately a definition for ψ_2 in terms of φ_1, e.g. $\psi_2 \leftrightarrow \varphi_1$, and vice versa, e.g. $\varphi_1 \leftrightarrow \psi_2$ such that $\{\varphi_1\}$ is synonymous to $\{\psi_2\}$. But then the logical consequence relation would be language dependent. The problem hinges here of course on the multiple definition of φ_1 and is avoided by demanding an overall definitional extension of all terms in the relata of the first relation or operation with the help of terms in the relata of the second relation or operation and vice versa. An easy and unproblematic example of a language dependent operation is the operation of counting the axioms of theories.

In order to talk not only about language dependent relations and operations, but also about language dependent theories, we may say that a theory is language dependent if it contains a relation or operation that is language dependent with respect to the theory:

Definition 6.3. Φ is language dependent iff there is an n-ary relation R^n or operation f^n of Φ such that R^n or f^n of Φ is language dependent.

Of course, in evaluating a theory as language dependent one has to take into account that many theories contain language dependent operations as, e.g., counting procedures, which are not at the heart of the theories. Take, e.g., a theory of simplicity which measures the degree of simplicity of a model (polynomial) by the degree of the polynomial. Of course the calculation of the degree of a polynomial is language dependent insofar the degree of, e.g., $y = f_1(x)^2$ is 2, whereas the degree of, e.g., $y = f_2(x)$ is 1, although by defining $f_2(x) = f_1(x)^2$ and $f_1(x) = f_2(x)^{-2}$ all theories containing exactly one of the polynomials are synonymous. Yet the problem of such a theory of simplicity hinges not on the calculation of the degree of a polynomial, but on the identification of the degree of simplicity with this degree. And that the measure of simplicity defined in such a way is language dependent makes such a theory a problematic one, not that it also contains counting procedures etc. So, strictly speaking, one should modify definition 6.3 by adding a condition like R^n and f^n are central in the theory Φ or their characterisations are the by Φ intended explications etc. But of course such an additional condition would make the whole clarification quite pragmatic and unclear again and so we just formulate the constraint of language independence as the constraint for a theory to be language independent in the above defined way, of course bearing in mind that we apply the definitions above only to the most relevant and central notions of the theory.

As we have motivated and clarified the constraint of language independence, we will now move on with a very general discussion of some widely accepted or investigated theories that do not satisfy this constraint: It can be shown that very many common, interesting and also fundamental and widely accepted notions of different areas of science and philosophy are language dependent. In what follows we are going to indicate this fact by discussing language dependent notions of some main areas of philosophy. Since many of these notions are quite general and cross-discipline wide in use, the subsumption of the following discussions under specific headings is more due to aesthetical than to intrinsic systematic reasons.

An example of language dependence in philosophy of science. We can rephrase Goodman's *new riddle* as discussed in section 6.1 as a problem of language dependency: The argument shows, e.g., that at least some common notions of confirmation are language dependent. Let $conf$ be a confirmation measure for theories that satisfies a principle of enumerative induction. Then instances of a general statement confirm other instances of the statement—at least to a higher degree than negative instances are confirmed. So, let us assume that for such a confirmation measure it holds:

$$\begin{gathered} conf(\{\varphi[a_n]\}, \{\varphi[a_1] \mathbin{\&} \cdots \mathbin{\&} \varphi[a_{n-1}]\}) > \\ conf(\{\neg\varphi[a_n]\}, \{\varphi[a_1] \mathbin{\&} \cdots \mathbin{\&} \varphi[a_{n-1}]\}) \end{gathered}$$

Now, take as example the Goodmanian emerald case: The given data can be summarised as a theory claiming that all emeralds observed until and including time n were *green*:

$$\mathcal{T}_1 = \{Pa_1, \ldots, Pa_{n-1}\}$$

Of course the question arises whether the emerald observed at time n will be also green or blue, i.e. not green? Let us take Φ_1 to be an affirming, Ψ_1 to be a negating theory:

$$\begin{aligned} \Phi_1 &= \{Pa_n\} \\ \Psi_1 &= \{\neg Pa_n\} \end{aligned}$$

By the principle of confirmatorial induction it holds:

$$conf(\Phi_1, \mathcal{T}_1) > conf(\Psi_1, \mathcal{T}_1)$$

Now, let us switch the language by defining a *grue* predicate Q as in the discussion before:

$$Qx \leftrightarrow (Px \leftrightarrow (x = a_1 \vee \cdots \vee x = a_{n-1}))$$

We can provide a definition of P by help of Q simply by swapping P and Q in this definition. Then the translation of the data reads as: All emeralds observed until and including time $n-1$ are *grue*, i.e.:

$$\mathcal{T}_2 = \{Qa_1, \ldots, Qa_{n-1}\}$$

And the theory stating that the emerald observed at n will be green, Φ_1, translates as: The emerald observed at n will be not *grue* (but *bleen*), i.e.:

$$\Phi_2 = \{\neg Qa_n\}$$

The translation of the competing theory stating that the emerald observed at n will be not green reads as: The emerald observed at n will be *grue*:

$$\Psi_2 = \{Qa_n\}$$

By applying the confirmatorial principle of induction it holds:

$$conf(\Phi_2, \mathcal{T}_2) < conf(\Psi_2, \mathcal{T}_2)$$

And so, someone claiming that the emerald observed at n will be not green instead of green is confirmed above its competitor if she switches the basic language. Since Φ_1 and Φ_2 as well as Ψ_1 and Ψ_2 are synonymous theories, it turns out that a confirmation measure that satisfies a confirmatorial principle of induction is language dependent. However, confirmation $conf$ and epistemic justification J in general are not the only notions which turn out to be language dependent. As we will show now, there are other very important notions which have this problem.

An example of language dependence in logic (in a wide sense). Take, e.g., the qualitative notion of verisimilitude introduced by Popper (1972, p.52). According to this account a theory Φ is closer to the truth than another theory Ψ, if Φ's truth-content, but not its falsity-content exceeds that of Ψ or if the falsity-content of Ψ, but not its truth-content exceeds that of Φ. (Note that this notion of verisimilitude is only logical in a similar sense as the notion of probability introduced by Carnap is logical.) As was shown independently by Tichý (1974) and Miller (1974) this theory of verisimilitude faces the problem that out of two false theories never one can be shown to be closer to the truth than the other. So, Popper's theory was modified by several authors: Oddie (2013) distinguishes three types of approaches: the *content approaches* which try to overcome the problem by specifying different contents (see, e.g., Oddie 2013); the *consequence approaches* which employ different consequence relations for explicating Popper's idea (see, e.g., Schurz and Weingartner 1987, who use a relevance consequence relation; and Schurz and Weingartner 2010); and the *likeness approaches* which employ distance measures for spelling out the notion of *likeness to the truth* (see, e.g., Tichý 1976; and Niiniluoto 1987). For simplicity reasons we concentrate here on a simplified form of the latter—namely a counting approach of verisimilitude. The idea of the counting approach of verisimilitude is to count the number of basic truths of a theory, the so-called true literals of the theories, and then to compare this number with the number of true literals of rival theories (although we wont employ the full terminology, here is how Niiniluoto (1987) unfolds it: a literal is an un-/negated atomic formula $\pm\varphi_1, \ldots$; a constituent is a state description of a possible world by help of a conjunction of literals $\pm\varphi_1 \& \pm \psi_1 \& \pm \chi_1 \& \cdots$; a theory consists in a set of disjunction of constituents). According to a simplified counting approach of verisimilitude a theory is sayd to be closer to the truth than another one if its number of true literals exceeds that of the other one (the here underlying distance measure is the so-called *Hamming distance* which takes as the distance of two constituents the number of diverging literals; for an overview of more common and also more sophisticated measures see Schurz and Weingartner 2010, sect.3). Let us illustrate this by help of an example. Let $\Delta(\Phi, \Psi)= |\Phi| - |\Phi \cap \Psi|$ be a distance measure that counts the distance of two sets of literals (in our toy example we use single constituents represented by a set of literals for a theory only). Let the truth be:

$$\mathcal{T}_1 = \{\varphi_1, \psi_1, \chi_1\}$$

And let two competing theories Φ_1 and Ψ_1 claim:

$$\begin{aligned} \Phi_1 &= \{\neg\varphi_1, \psi_1, \chi_1\} \\ \Psi_1 &= \{\neg\varphi_1, \neg\Psi_1, \neg\chi_1\} \end{aligned}$$

Then the competing theories' distance from the truth calculates as follows: $\Delta(\mathcal{T}_1, \Phi_1) = 1$ and $\Delta(\mathcal{T}_1, \Psi_1) = 3$. So Φ_1 is closer to the truth than Ψ_1:

$$\Delta(\mathcal{T}_1, \Phi_1) < \Delta(\mathcal{T}_1, \Psi_1)$$

Now, as was already put forward by Miller (1974, sect.6), the counting approach of verisimilitude is language dependent. For a demonstration of this fact let us switch the language from index-1-expressions to index-2-expressions by defining:

$$\begin{aligned} \varphi_2 &\leftrightarrow \varphi_1 \\ \psi_2 &\leftrightarrow (\varphi_1 \leftrightarrow \psi_1) \\ \chi_2 &\leftrightarrow (\varphi_1 \leftrightarrow \chi_1) \end{aligned}$$

It is also possible to give analogous definitions for $\varphi_1, \psi_1, \chi_1$ in terms of $\varphi_2, \psi_2, \chi_2$ by swapping φ_1 and φ_2, ψ_1 and ψ_2, and χ_1 and χ_2 respectively in these definitions. Then the translation of the truth and the two competing theories reads as:

$$\begin{aligned} \mathcal{T}_2 &= \{\varphi_2, \psi_2, \chi_2\} \\ \Phi_2 &= \{\neg\varphi_2, \neg\psi_2, \neg\chi_2\} \\ \Psi_2 &= \{\neg\varphi_2, \psi_2, \chi_2\} \end{aligned}$$

The competing theories' distance from the truth calculates now: $\Delta(\mathcal{T}_2, \Phi_2) = 3$ and $\Delta(\mathcal{T}_2, \Psi_2) = 1$. So Ψ_2 is closer to the truth than Φ_2:

$$\Delta(\mathcal{T}_2, \Phi_2) > \Delta(\mathcal{T}_2, \Psi_2)$$

Since Φ_1 and Φ_2 as well as Ψ_1 and Ψ_2 are synonymous it follows that the simple distance measure Δ is language dependent and by this also the logical theory of verisimilitude in the counting approach is language dependent.

An example of language dependence in epistemology. Impossibility results in social choice theory and social epistemology show that no aggregation of individual judgements can be universal in the sense that the whole algebra of a language and all probabilistically possible judgements are covered, anonymous in the sense that a permutation of the individual judgements leads to identical results, systematic in the sense that an aggregated judgement about a statement is achieved by individual judgements about this statement only and not depending on individual judgements about other statements, and logically and probabilistically consistent (for an overview see Pivato 2008; for a prominent impossibility result of social epistemology see List and Pettit 2002). We will discuss this problem in detail in chapter 11. Since in cases of majority voting all ingredients of an aggregation, namely individually coherent judgements and the majority voting procedure itself, seem to be plausible—take as an example the following judgements of the member of a jury (see Kornhauser 1992, p.454):

Claims:	φ	$(\varphi \rightarrow \psi)$	ψ
Judge$_1$	0	0	0
Judge$_2$	1	0	1
Judge$_3$	1	1	1

And since the output of such an aggregation may not be plausible, as, e.g., a majoritarian aggregated judgement turns out to be incoherent:

Claims:	φ	$(\varphi \rightarrow \psi)$	ψ
Court	1	1	0

such a scenario was labelled as a 'paradox' (see Kornhauser 1992, p.454). Since the impossibility result mentioned above shows that the formation of an implausible output out of a plausible or coherent input does not hinge specifically on the majority aggregation method, but on any aggregation method satisfying the above mentioned constraints, the situation was more generally labelled as a 'dilemma' (see List and Pettit 2002, pp.89ff).

In order to cope with the impossibility results in an adequate way, different proposals were made to weaken the conditions for judgement aggregation. One of the most prominent proposals is a weakening of the systematicity condition by taking into account not only the individual judgements on a proposition, but also the individual judgements on logically related propositions. So, e.g., so-called premise-based aggregation functions were proposed which violate only the systematicity constraint by outweighing an aggregation result of premisses against an aggregation result of conclusions in case of a conflict (see, e.g., List and Pettit 2011, chpt.4.2; Hartmann and Sprenger 2012). Premise-based majority voting, e.g., would turn the court's decision in the example above into a coherent one:

Claims:	φ	$(\varphi \rightarrow \psi)$	ψ
Court	1	1	1

But again, one can show that such a premise-based majoritarian aggregation is language dependent (see Cariani, Pauly, and Snyder 2008): Let $aggr$ be such a premise-based majoritarian aggregation rule. Now, take three individual judgements or theories:

$$\begin{aligned}\Phi_1 &= \{\varphi_1, \psi_1, (\varphi_1 \;\&\; \psi_1)\}\\ \Psi_1 &= \{\varphi_1, \neg\psi_1, \neg(\varphi_1 \;\&\; \psi_1)\}\\ \mathcal{X}_1 &= \{\neg\varphi_1, \psi_1, \neg(\varphi_1 \;\&\; \psi_1)\}\end{aligned}$$

Let us assume that φ_1 and ψ_1 count as premisses and $\varphi_1 \;\&\; \psi_1$ as conclusion. Then, applying a simple majority rule would lead to the logical inconsistent judgements:

$$\text{Majoritarian aggregation: } \{\varphi_1, \psi_1, \neg(\varphi_1 \;\&\; \psi_1)\}$$

And so $aggr$ favours the premisses' outcome ending with an aggregation of these judgements as follows:

$$aggr(\Phi_1, \Psi_1, \mathcal{X}_1) = \{\varphi_1, \psi_1, (\varphi_1 \;\&\; \psi_1)\}$$

Now, again, let us switch from index-1 expressions to index-2 expressions by defining:

$$\begin{aligned}\varphi_2 &\leftrightarrow \varphi_1\\ \psi_2 &\leftrightarrow (\varphi_1 \leftrightarrow \psi_1)\end{aligned}$$

By swapping φ_2 and φ_1, and ψ_2 and ψ_1 respectively in these definitions one assures intertranslatability. Expressing the judgements in this new language, the translations are as follows:

$$\begin{aligned}\Phi_2 &= \{\varphi_2, \psi_2, (\varphi_2 \;\&\; \psi_2)\}\\ \Psi_2 &= \{\varphi_2, \neg\psi_2, \neg(\varphi_2 \;\&\; \psi_2)\}\\ \mathcal{X}_2 &= \{\neg\varphi_2, \neg\psi_2, \neg(\varphi_2 \;\&\; \psi_2)\}\end{aligned}$$

Now, applying the simple majority rule leads to a consistent judgement and so the aggregation turns out to be:

$$aggr(\Phi_2, \Psi_2, \mathcal{X}_2) = \{\varphi_2, \neg\psi_2, \neg(\varphi_2 \;\&\; \psi_2)\}$$

But note that the translation of $aggr(\Phi_1, \Psi_1, \mathcal{X}_1) = \{\varphi_1, \psi_1, (\varphi_1 \;\&\; \psi_1)\}$ is: $\{\varphi_2, \psi_2, (\varphi_2 \;\&\; \psi_2)\}$. So, although Φ_1, Φ_2, Ψ_1, Ψ_2, and $\mathcal{X}_1, \mathcal{X}_2$ are synonymous theories it holds:

$$aggr(\Phi_1, \Psi_1, \mathcal{X}_1) \neq aggr(\Phi_2, \Psi_2, \mathcal{X}_2)$$

This means that also the premise-based majoritarian aggregation procedure $aggr$ is language dependent in the weakened theory of social opinion pooling and by this the theory itself is language dependent.

An example of language dependence in ethics. A very similar problem arises in ethics when one bases her theory on preference aggregation. Take, e.g., a line of argumentation in the debate on value-free science. One quite common proposal in this debate is to consider the context of justification of science as being totally free of non-epistemic values and that a scientist *qua* scientist should provide for an application of her results just information about the efficiency of means-ends relations (see Schurz 1997, pp.239ff as well as 11.3 and 11.4). So, according to this proposal, e.g., in order to apply scientific results, scientific theories provide information about the probability of achieving some ends by some means, but which ends and means are chosen depends not only on these probabilities, but also on the utilities of the ends and means, which are achieved by aggregating the preferences of all by a policy influenced persons. Since preference aggregation is structurally similar to judgement aggregation, also the problems of the latter hold for the former (see List and Pettit 2011, pp.46f): Also such a theory turns out to be language dependent. Take, e.g., the aggregations $aggr(\Phi_1, \Psi_1, \mathcal{X}_1)$ and $aggr(\Phi_2, \Psi_2, \mathcal{X}_2)$ from above, so in terms of an index-1 language people prefer both φ_1 and ψ_1, but in terms of an index-2 language people have a preference only for φ_2, but not for ψ_2, and let us assume that scientific tests show that ψ_1 is a good means to achieve ends φ_1 ($Pr(\varphi_1|\psi_1)$ is quite high; let us assume that scientific tests also show $Pr(\varphi_2|\psi_2)$ to be quite high). Then, since the means is only preferred in the index-1 language (ψ_1), but not in the index-2 language (ψ_2), the advice to use the means in order to achieve the ends (a high enough expected utility) will be only given in the index-1, but not in the index-2 language. And so a decision procedure based on such a preference aggregation turns out to be language dependent too.

An example of language dependence in metaphysics. In recent times it was shown that also some very common notions of metaphysics, e.g., the Bayesian notion of causality, are language dependent (see Spirtes 2009). In order to illustrate this fact, we will provide a short discussion in terms of the so-called SGS-algorithm of Spirtes, Glymour, and Scheines (2000, p.82). Let us assume that we have grasped some statistical data about the weather at some specific area—let us consider here only a very simple statistics providing only very rough information about the weather conditions for each day of a year. So, the statistics looks something like the following table:

Day	1.	2.	3.	4.	5.	…	361.	362.	363.	364.	365.
Cold (φ_1)	1	1	0	1	0	…	1	1	1	1	1
Rainy (ψ_1)	0	0	0	0	1	…	1	1	1	1	1
Snowy (χ_1)	0	0	0	1	0	…	0	0	1	1	1

Now, assume that at the imagined place a little bit less than one third of the year's days were cold and that the same percentage holds also for rainy and snowy days. Furthermore, let us assume that it was cold, when it was rainy and vice versa on 30% of the days and that only on 30% of these days it was also cold. Finally, let our statistical records claim that only on 9% of the days it was rainy and snowy. So our statistical data can be summarised as follows:

$$\begin{aligned}
Pr(\varphi_1) = Pr(\psi_1) = Pr(\chi_1) &= 0.30 \\
Pr(\varphi_1 \leftrightarrow \psi_1) &= 0.30 \\
Pr((\varphi_1 \leftrightarrow \psi_1) \,\&\, \varphi_1) &= 0.09 \\
Pr(\psi_1 \,\&\, \chi_1) &= 0.09
\end{aligned}$$

From this data it follows, that $Pr(\psi_1|\chi_1) = Pr(\psi_1)$ and $Pr((\varphi_1 \leftrightarrow \psi_1)|\varphi_1) = Pr((\varphi_1 \leftrightarrow \psi_1))$. So rainyness and snowyness are statistically independent as well as coldness is statistically independent of the co-occurrence of coldness and rainyness. Let us assume that these are the only relevant independences in our sample. Now, imagine that within a simple meteorological investigation the causal relation between these three factors should be brought to the light. For this purpose we can, e.g., apply an algorithm of the Bayesian theory of causality to our statistical data. Let us do so by applying the SGS-algorithm stepwise:

(A) First, we list all relevant independences for our variables and then draw a complete undirected graph with our variables. We use as binary causal variables $X_{\varphi_1}, X_{\psi_1}, X_{\chi_1}$. The only relevant independence for our simple example is $Ind_{Pr}(X_{\psi_1}, X_{\chi_1})$. We end up with an undirected graph as shown in figure 6.1.

(B) Now, in a second step, we consider our variables pairwise and check whether they are independent conditional on any (possibly empty) subset of the remaining variables. If so, then we remove the connection between them. Since the only independence is between X_{ψ_1} and X_{χ_1} conditional on $\emptyset$, we end up with the graph shown in figure 6.2.

(C) Now we try to figure out the direction of the causal relation. For this purpose we consider in a next step all connected parts of the graph of the form ○—○—○ (NB: exclude cases of such a form where the edges are also connected). Whenever the edges of such a part are independent conditional on every subset of variables containing at least the variable in between them, then we have figured out a common effect of the form: $\circ \longrightarrow \circ \longleftarrow \circ$. Since the only part of our graph of the form ○—○—○ is the graph itself and since the only variable remaining when we extract the edge-variables is X_{φ_1}, we only have to check whether $Ind_{Pr}(X_{\psi_1}, X_{\chi_1}|X_{\varphi_1})$ holds or not and since this is not the case, we end up with the graph shown in figure 6.3.

(D) In the next step we have to repeat the following directives until no more edges can be oriented:

- If $A \longrightarrow B$, B—C and neither A—C nor $\circ\!\longrightarrow B$, then orient B—C as $B \longrightarrow C$!
- If there is a directed path from A to B, and an edge between A and B, then orient A—B as $A \longrightarrow B$!

Since neither of the conditions of these steps is satisfied, our graph remains unchanged (see figure 6.3).

(E) And finally we simply "check" the so-called *Causal Markov Assumption* (all variables are independent of their non-effects conditional on their immediate causes) and the *Causal Faithfulness Assumption* (there are exactly those independences that are entailed by the Causal Markov Condition). Since X_{φ_1} has no non-effect and the only non-effects of X_{ψ_1} and X_{χ_1} are X_{χ_1} and X_{ψ_1} respectively, and since X_{ψ_1} as well as X_{χ_1} have no immediate causes, the Causal Markov Assumption entails exactly $Ind_{Pr}(X_{\psi_1}, X_{\chi_1})$. Since this is also the only independence in our example, also the Causal Faithfulness Assumption is satisfied. So we are done.

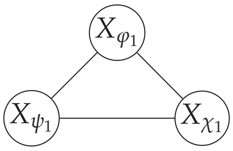

Figure 6.1: Performing the SGS-algorithm: (A)

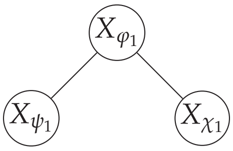

Figure 6.2: Performing the SGS-algorithm: (B)

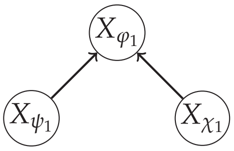

Figure 6.3: Performing the SGS-algorithm: (C)–(E)

An application of the SGS-algorithm brings to the light that in our simple weather example both, rainyness and snowyness have as a common effect coldness.

After this small lesson in meteorology let us again come to a foreign index-2 language for which we again provide rules of intertranslation. Let us assume that we investigate coldness (φ_2) and the unsplitted consideration of the co-occurrence of coldness and rainyness (ψ_2) on the one hand and coldness and snowyness on the other (χ_2). In terms of the weather vocabulary above we can provide easy definitions for the index-2 expressions:

$$\begin{aligned} \varphi_2 &\leftrightarrow \varphi_1 \\ \psi_2 &\leftrightarrow (\varphi_1 \leftrightarrow \psi_1) \\ \chi_2 &\leftrightarrow (\varphi_1 \leftrightarrow \chi_1) \end{aligned}$$

If we interchange φ_1 with φ_2, ψ_1 with ψ_2 and χ_1 with χ_2, we end up with definitions of index-1 expressions by the help of index-2 expressions only, that are logically equivalent with these ones. Now, let us consider a little bit more complicated causal structure within this language. Let us ask if, and if so: what, causal relation holds between coldness (φ_2), the co-occurrence of coldness and the co-occurrence of coldness and rainyness (ψ_2) and the co-occurrence of coldness and the co-occurrence of coldness and snowyness (χ_2). We choose as binary causal variables $X_{\varphi_2}, X_{(\varphi_2 \leftrightarrow \psi_2)}, X_{(\varphi_2 \leftrightarrow \chi_2)}$.

The co-occurrence of coldness and the co-occurrence of coldness and rainyness is logically equivalent to the co-occurrence of coldness and rainyness ($(\varphi_2 \leftrightarrow \psi_2)$). We have seen that the co-occurrence of coldness and rainyness is independent of coldness. That is also the only relevant independence relation in our new language. So we know exactly $Ind_{Pr}(X_{(\varphi_2 \leftrightarrow \psi_2)}, X_{\varphi_2})$. Now, by performing the SGS-algorithm we end up with a causal graph as depicted in figure 6.4.

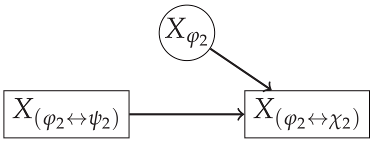
X_{φ_2}
$X_{(\varphi_2 \leftrightarrow \psi_2)}$
$X_{(\varphi_2 \leftrightarrow \chi_2)}$

Figure 6.4: Performing the SGS-algorithm with a new set of variables: Steps (A)–(E)

So coldness and the co-occurrence of coldness and rainyness cause the co-occurrence of coldness and snowyness. The latter is logically equivalent with the co-occurrence of coldness and the co-occurrence of coldness and snowyness. And since $\{\varphi_1\}$ is synonymous to $\{\varphi_2\}$ (represented by X_{φ_2} in the graph), $\{\psi_1\}$ is synonymous to $\{(\varphi_2 \leftrightarrow \psi_2)\}$ (represented by $X_{(\varphi_2 \leftrightarrow \psi_2)}$ in the graph) and $\{\chi_1\}$ is synonymous to $\{(\varphi_2 \leftrightarrow \chi_2)\}$ (represented by $X_{(\varphi_2 \leftrightarrow \chi_2)}$ in the graph), we end up with a different causal structure among synonymous theories. Hence, also the Bayesian notion of causality is language dependent.

So we have seen that many very common notions in logic, philosophy of science, epistemology, ethics, etc. do not satisfy the constraint of language independence (one can even show that the "logical and definitional" demarcation of analytic from synthetic statements is language dependent—for details see Feldbacher-Escamilla (2017a)). Since most of these notions are widely accepted and used, quite natural and in accordance with our everyday understanding of similar notions, it seems to be paradox that the correct application of these notions depends on the more or less arbitrary way we have chosen our language. As Miller (2006, p.215) summarises some critique on his constraint of language independence: It seems that "an argument which purports to show that the notions of accuracy, truthlikeness, structure, change, sameness of state, confirmation and disconfirmation, are all spurious [. . .] must harbour a defect somewhere."

As things stand, Goodman's riddle of induction concerns not only an encapsulated problem of epistemic justification J in form of a theory of induction or confirmation. Rather, it is an instance of the general problem of language dependency of many of our notions in logic, philosophy of science, ethics, metaphysics etc. We face the dilemma that we do not want to give up these notions, but also not the constraint of providing language independent justifications, inferences, assessments, etc. However, we cannot eat the cake and have it. So, what to do? Shall we keep our justified notion of induction and abandon the constraint of language independency, or shall we keep the latter and are once more in need of providing a justification of induction which cannot be turned to a justification of anti-induction? We will argue in the next section for the first solution and show that the cake (language independency) looks nice, but actually does not taste that well, for which reason one might neither want to keep nor eat it.

6.4 A Solution: Language Dependency of Language Dependency

We have seen that the constraint of language independence set up for the most important notions of theories of all areas of philosophy leads to quite implausible consequences. Since they do not preserve their structure among synonymous theories, the conventional part of choosing a primitive vocabulary when there are at least two equally well serving alternatives plays a crucial role in all these theories. And even worse, by many of them an ordering between two pairs of synonymous theories can be inverted through language conversion. So, e.g., a perfectly fine argument for an inductive inference can be turned into a perfectly fine argument for anti-inductive inference, just by switching from a 'green'/'blue' language to a 'grue'/'bleen' language. So, in applying these theories an alleged underdog can quite often beat her opponent just by switching the language. This sounds paradox.

We can explicate this paradox of allowing for definitional transformations by putting forward a constraint of language dependency with the following argument:

1. Some important relations R's and operations f's of different theories seem in general to be adequate with respect to the intended domain of application of these theories.

2. And also the constraint that the R's and f's should be language independent seems to be adequate.

3. Nevertheless it can be shown that the R's and f's are language dependent.

4. So, either you have to abandon the constraint of language independence or your notions (R's and f's)—at least in the form they are now.

But in general we do not think that to abandon our fundamental and quite fruitful notions is a good idea, nor do we think that language dependence is an unproblematic feature. What are possible ways out of this dilemma? Since premise 3 concerns just technical elaboration and is quite uncontroversial, there are only two ways left: First of all, one may bring oneself to doubt premise 1, as, e.g., Miller (2006, chpt.11) did: If one takes the constraint of language independence serious and if it turns out that at first glance adequate seeming notions fall short of satisfying this constraint, then one should abandon these notions or try to fix the problem by modifying them until they satisfy this constraint. Second, one may doubt the constraint of language dependence. A very systematic and summarising overview of this line of argumentation can be also found in (Miller 2006, chpt.11). In principle the argumentation against premise 2 can be quite manifold: (i) One might accept that a theory's outcome essentially depends on the language chosen for the theory and by this accept some kind of relativism: E.g., whether a theory is closer to the truth against another one depends in the end also on the languages we use to formulate the theories in (for representatives and discussion see Miller 2006, chpt.11, sect.5). (ii) One might accept that a theory's outcome depends on the language chosen, but one may not accept that all languages serve the endeavour of science equally well. So theories need not to preserve their structure among synonymous theories, but only among synonymous theories formulated in acceptable languages (see Miller 2006, chpt.11, sect.3). (iii) One might accept some constraint of language independence, but not that one presented here (see Miller 2006, chpt.11, sect.4).

In this section we argue against premise 2 in the line of (i). We do so by showing an implausible consequence of this plausible constraint of language independence: One can show that the property of language independence is itself language dependent. So, putting forward a constraint of language independence is self defeating. In order to do so one just needs to prove that there are two synonymous theories Φ_1 and Φ_2 and that one of them is language dependent whereas the other is not. Take the following example formulated in a language of first-order logic: We assume that a_1 and b_1, and a_2 and b_2 respectively are synonymous theories! Furthermore we assume that Φ_1 and Φ_2 are theories (about theories) as follows:

- $\Phi_1 = \{a_1 \neq b_1 \;\&\; \forall x P_1 x\}$

- $\Phi_2 = \{a_2 \neq b_2 \;\&\; \forall x P_2 x \leftrightarrow x \neq a_2\}$

One may take Φ_1 to be a naïve sceptical theory about theories claiming that all theories are false. And Φ_2 could be seen as a little bit more reflecting sceptical theory that states, e.g., that all theories are false except the theory itself (a_2). Now, since a_1, b_1 and a_2, b_2 are assumed to be synonymous, and since $\Phi_2 \vdash (P_2 b_2 \,\&\, \neg P_2 a_2)$ it follows that P_2 of Φ_2 valuates synonymous theories differently and so by definition 6.2 P_2 of Φ_2 is language dependent. On the other hand, since P_1 of Φ_1 valuates all—also synonymous—theories on a par, namely to be false, P_1 of Φ_1 is not language dependent. As Φ_2 contains a language dependent property and Φ_1 does not we can say that Φ_2 is a language dependent sceptical theory whereas Φ_1 is a language independent one.

Now, one can easily show that Φ_1 and Φ_2 are synonymous theories. Take, e.g., the following sets of definitions:

$$\begin{aligned} \mathcal{D}_{\Phi_2}: \quad P_1 x \leftrightarrow ((x = a_2 \rightarrow \neg P_2 x) \,\&\, (x \neq a_2 \rightarrow P_2 x)) \\ a_1 = a_2, b_1 = b_2 \end{aligned}$$

And:

$$\begin{aligned} \mathcal{D}_{\Phi_1}: \quad P_2 x \leftrightarrow ((x = a_1 \rightarrow \neg P_1 x) \,\&\, (x \neq a_1 \rightarrow P_1 x)) \\ a_2 = a_1, b_2 = b_1 \end{aligned}$$

Then it holds that Φ_1 and Φ_2 translatable into each other by help of these definitions: $\Phi_1 \cup \mathcal{D}_{\Phi_2} \dashv\vdash \Phi_2 \cup \mathcal{D}_{\Phi_1}$. It even holds that these definitions are logically equivalent: $\mathcal{D}_{\Phi_2} \dashv\vdash \mathcal{D}_{\Phi_1}$. So Φ_1 and Φ_2 are intertranslatable, i.e. synonymous.

Since Φ_1 is not language dependent, but Φ_2 is, and since Φ_1 and Φ_2 are synonymous it follows that the property of language dependence does not preserve its structure among synonymous theories and is for this reason itself language dependent. So, the problem that sometimes a bad valuation can be turned into a good one just by switching the language can sometimes be overcome by switching the language again.

Note that in the example the two theories are inconsistent given the definitions. So, one could, e.g., strengthen the notion of *synonymy* by demanding consistency of the theories given the definitions. However, although we were not able to quickly find a counterexample for such a strengthened constraint, we conjecture that one can find one by employing a bit more structure than we did here. Furthermore, note that one might also reverse the direction of our argument as follows (thanks to Gerhard Schurz for pointing out this possible objection): Since Φ_1 and Φ_2 are synonymous and since Φ_2 is language dependent, also Φ_1 is. But such an argument would presuppose a notion of *language dependency* granting a principle of the following form: If x is language dependent and x is synonymous with y, then also y is language dependent. However, we do not see how such a principle could be constructed out of the notion of *language dependency* as introduced above. For this reason we think that the argument shows that the property of language independence (at least as it is defined above) is itself not language independent, and hence putting forward language independence as a constraint is self-refuting. Coming back to the problem of justifying anti-induction Goodman style, we think that premise 4 (p.183) of the argument is inadequate: Claiming that justification needs to be preserved among definitional transformations means that justification needs to be language independent. However, even the notion of *language independency* itself falls short of this constraint. How then can our notion of *justification* satisfy it? And even if it were to satisfy it in one language, it might fail to satisfy it in another one. So, in our understanding the justification of anti-induction via definitional transformations fails, because *justification* is not and need not be language independent, i.e. invariant under such transformation. Our solution to the new riddle of induction is illustrated in the schema of table 6.4.

$$\begin{tikzcd}
s \arrow[r, Rightarrow] \arrow[d, Leftrightarrow, "\rotatebox{90}{df}"'] & f_{mi} \arrow[r, Rightarrow] \arrow[d, Leftrightarrow, "\rotatebox{90}{df}"'] & \text{optimality} \arrow[r, Rightarrow] & J \arrow[d, Leftrightarrow, "/" marking] \\
\rotatebox{180}{s} \arrow[r, Rightarrow] & f_{m\rotatebox{180}{i}} & & \rotatebox{180}{J}
\end{tikzcd}$$

Table 6.4: A justification of meta-anti-induction $f_{m\rotatebox{180}{$i$}}$ by help of meta-induction f_{mi} fails, since justification is not preserved among definitional transformations.

Chapter 7

Abduction and Optimisation

In this chapter two forms of abduction are distinguished, selective and creative abduction. An exact characterisation of selective abduction in terms of accuracy and simplicity is provided. Afterwards, the epistemic merits of simple explanations and predictions are discussed. Then an exact characterisation of creative abduction as a form of inference to a common cause explanation and prediction is provided. Finally, it is outlined how the theory of meta-induction can be also employed for justifying both kinds of abductive inferences.

Up to now we have argued that the theory of meta-induction allows for justifying inductive inferences based on their past successes. However, there is also another inference method which is widely-used in science and which is also in need of epistemic justification, namely the method of abduction. The question of justifying this method will be addressed in the present chapter.

Charles S. Peirce was the first to describe *abductive inferences* as a topic of philosophy of science and logic in the broad sense. He characterised such an inference schematically as follows (see Peirce 1994):

1. The surprising fact, E, is observed;
2. But if H were true, E would be a matter of course;
3. Hence, there is reason to suspect that H is true.

More generally, we can distinguish two kinds of abductive inferences: those generating new hypotheses and those aiming at determining the best hypothesis from a set of available candidates. Abductive inferences of the former kind are sometimes called *creative abductions*, and those of the latter kind *selective abductions* (see, e.g. Magnani 2000; Schurz 2008a). Selective abduction is often subsumed under the term *inference to the best explanation* and most of the philosophical literature on abduction focuses on this form of abductive inferences (see, e.g., Lipton 2004; Niiniluoto 1999; Williamson 2016). However, there is also an increasing interest in creative abduction (see Douven 2018) which is intended as an inference method for generating hypotheses featuring new theoretical concepts on the basis of empirical phenomena.

At least at first glance, it seems clear that combining creative and selective abduction allows for a powerful inference tool, as in principle any combination of several inference methods does. Starting from empirical data, creative abduction might be used for *inferring* whole theories which can then, in turn, be used as input for selective abduction in order to *infer* the best one. Now, clearly there is the question of how to characterise these abductive methods exactly. Connected to this is also the question of how to justify them as viable inference methods. Both problems will be addressed here.

We will proceed as follows: In section 7.1 we characterise *selective abduction* roughly as an inference to the *best* explanation, where *best* is understood in terms of accuracy and simplicity. Now, the epistemic value of accurate explanations and predictions is clear, but that of simplicity needs to be explored further. This is done in section 7.2, where an information theoretical argument in favour of simplicity is discussed and used for fleshing out the notion of a *selective abductive inference* further. Afterwards, in section 7.3, we characterise creative abduction in detail and evaluate its ability to achieve simpler explanations and predictions. Finally, in section 7.4, we use the theory of meta-induction to argue for a positive answer to the question whether abduction can be justified or not.

7.1 Selective Abduction

As we have outlined in the first chapter of this part, there are three major types of inference used in science and philosophy: deduction, induction, and abduction. Deductive inferences are truth preserving. Inductive inferences are not truth preserving, but have conclusions containing predicates that occur already in the premisses. Finally, abduction is formally characterised as a non-deductive inference with a conclusion containing also predicates that do not occur already in the premisses. So, e.g., $\{\forall x R(x)\} \vdash R(c)$ is a deductive inference. $\{R(c_1), \ldots, R(c_n)\} \mid\!\sim \forall x R(x)$ is an inductive inference. And the inference from

$$\begin{gathered}\{\exists_1^1 x R(x, t_1), \exists_1^1 x W(x, t_1), \forall x (R(x, t_2) \& \neg W(x, t_2)),\\ \exists_n^n x R(x, t_3), \exists_{n/4}^{n/4} x W(x, t_3)\}\end{gathered}$$

to

$$\exists_1^1 x E(x, t_1) \& \exists_1^1 x D(x, t_1) \& \neg \exists x M(x, t_1)$$

is an (abbreviation of an) abductive one, since E, D, and M do not occur in the premise set ('$\exists^n_n$' stands for 'there are exactly n'). Abductive inferences play a major role in natural sciences as they are widely used for theory construction. Simplified speaking, by abduction one can infer from empirically accessible data theoretical hypotheses that allow for a more or less simple explanation of the data. Prominent is the example of Gregor Mendel inferring from phenotypic properties of plants (e.g. colours red R and white W) laws of inheritance (e.g. the inheritance of recessive E, dominant D, and mixed M traits). By this he was able to explain, e.g., why and how much white plants are to be expected in the third generation, although this phenotypic property seemed to have died out in the second generation. The schema of this prototypical abductive inference is depicted in figure 7.1.

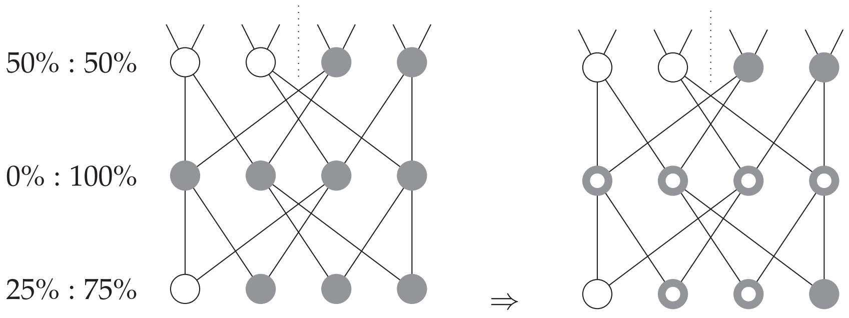

Figure 7.1: A prototypical abductive inference: Gregor Mendel's famous laws of inheritance: In the 1850s and 60s, Mendel cultivated and tested about 5.000 pea plants and performed hybridisation experiments. Mendel inferred from regularities about R, W (red, white colour), laws about E, D, M (recessive (white), dominant (red), mixed traits (red and white)). The data is presented on the left side. The inferred structure on the right. The edges represent inheritance. The underlying theoretical structure was simple and allowed for an empirically adequate explanation of Mendel's data.

Now, are there any further characteristics of abductive inferences? Besides the formal constraint of introducing new (theoretical) vocabulary, materially speaking characteristic for abduction is its validation of explanations. So, usually an abductive inference has no single statement as a conclusion, but laws and regularities that can be used in an explanation or that even form a whole theory. In the case of Mendel's abductive inference, given the premise set above the laws and regularities of a validated explanation might consist of assumptions about the initial traits as presented above (one dominant and one recessive) as well as probabilistic reasoning based on assumptions about the average number of descendants of each possible pair per each generation.

What are the constraints for validating such an explanation? Abduction in the sense of an *inference to the best explanation* (see, e.g., Lipton 2004) is usually supposed to maximise the data's plausibility and the hypotheses' simplicity. Regarding the former, the parameter consists in the probability of the premise' P (also: the *explanandum*'s) in the light of laws and regularities used in the explanation (the conclusion C or also: the *explanans*). The simplicity constraint is considered to be necessary in order to rule out ad hoc explanations. For, if one takes, e.g., scientists' conditional degrees of belief Pr of the explanandum P in the light of the explanans C as a measure for plausibility: $Pr(P|C)$, which is the likelihood of P given C, then it is clear that choosing a C such that $C \vdash P$ maximises the explanandum's plausibility in the light of the explanans. In the simplest case one might set ad hoc: $C = P$. However, what we aim at are not ad hoc explanations that might be even trivial, but universal explanations. Since ad hoc explanations usually turn out to become more and more complex with an increased number of data, abductive validation of an explanation hinges not only on $Pr(P|C)$, but also on C's simplicity. If we assume that there is some way of measuring C's complexity via a non-negative function $c(C)$, then we can characterise the validation procedure of an abductive inference as trying to maximise $Pr(P|C)$ on the one side, and minimise $c(C)$ on the other.

Several remarks are in place: First, in order to remain applicable, in this method the aim of maximising $Pr(P|C)$ and minimising $c(C)$ is to be understood not in absolute terms, but in relative ones. We might aim at $Pr(P|C) = 1$ and $c(C) = 0$, but we will almost never achieve this goal. In particular, it is presupposed that we exclude trivial abductive inferences to P itself ($Pr(P|P) = 1$ is maximal). As we have mentioned above, with an increased number of data P to be plausibly explained by C usually also the complexity of C increases. And on the other hand, reducing the complexity of C usually leads to generalisations of C that are not in full agreement with P, for which reason $Pr(P|C)$ decreases. Since these two measures are intertwined in many applications, often finding a C such that $Pr(P|C) = 1$ and $c(C) = 0$ is not achievable. This was highlighted, e.g., also by Popper, who claimed that the aim of increasing $Pr(P|C)$ "inadvertently but necessarily, implies the unacceptable rule: always use the theory which is the most *ad hoc*, i.e. which transcends the available evidence as little as possible [i.e. which sets $C = P$]" (see Popper 2002a, p.61). However, what is clearly achievable is a comparative task: Assume that the only available *potential explanantia* are $C_1, \ldots, C_n$. Then it holds:

$$\begin{aligned} &\text{If there is a } i \in \{1, \ldots, n\} \text{ such that for all } j \in \{1, \ldots, n\} \setminus \{i\}:\\ &\qquad c(C_i) \leq c(C_j) \;\&\; Pr(P|C_i) > Pr(P|C_j)\\ &\qquad \text{or}\\ &\qquad c(C_i) < c(C_j) \;\&\; Pr(P|C_i) \geq Pr(P|C_j),\\ &\qquad \text{then infer from } P \text{ by abduction } C_i \end{aligned} \tag{Abd}$$

(Abd) demands that in case there is an explanans C_i which plausibilises P better, but not at cost of being more complex, or which is simpler, but still not at cost of less plausibilising P than all the other possible explanantia, that in such a case C_i is to be inferred from P. If one generalises this comparative validation to the set of all potential explanantia *one has thought of* (see Williamson 2016, p.267), then one might regain an absolute phrasing of abductive inferences that is still applicable.

Second, $Pr(P|C)$ and $c(C)$ can be balanced in several ways. One might consider, e.g., a combination of the form $(1 - Pr(P|C)) \cdot c(C)$ which is the product of the inverse of the likelihood and complexity that is to be minimised, but one might also think of maximising $Pr(P|C) - c(C)$. These possibilities of balancing lead to different inferences in at least some applications. However, what is important to note is that they still satisfy (Abd). This is also the minimal constraint we want to put forward for abduction and as long as a non-deductive inference rule introducing new vocabulary satisfies it, we think it is fine to call it an 'abductive' one. In the next section we will consider another way of balancing that also satisfies (Abd).

Third, the two parameters $c(C)$ and $Pr(P|C)$ are not sufficient for providing a fully adequate account of selective abduction. Usually, also the prior probabilities of the hypotheses used in an explanation are relevant. E.g., if $Pr(C_i)$ is very close to 0 and $Pr(C_j)$ is high, then one will still tend to opt for C_j instead of C_i, although $Pr(P|C_i)$ might be greater than $Pr(P|C_j)$. For simplicity reasons we restrict the application of (Abd) to cases with close prior probabilities of the alternative hypotheses $C_1, \ldots, C_n$. Also, as is pointed out in (Schurz 2008a), other theoretical virtues of C as, e.g., use-novelty, unification, etc. are typically considered to be relevant for abductive inferences. Again, we restrict the intended application of (Abd) to cases where these theoretical virtues are considered to be satisfied equally well. The reason for this strong restrictions is twofold. First, some of these further parameters might be reducible to the two we are proposing. So, e.g., regarding unification and use-novelty, Forster and Sober (1994) provide reduction strategies which might be cashed out be allowing for complex P and C (conjunctions of descriptions of phenomena and hypotheses). In principle, one might even think of reducing the prior probability of a hypothesis ($Pr(C)$) as relevant parameter via inversely relating it to the complexity measure $c(C)$ (this would be, e.g., along the lines of Solomonoff 1964). The second reason is that in this chapter we are only interested in an exemplary application of meta-induction to abductive inferences. For this purpose it suffices to show how the theory can be applied in case one scores not only according to accuracy, but also according to some theoretical value like simplicity/complexity. So, we aim at theorising only about a simple model of abduction which clearly has lots of limitations.

Now, one might wonder why $c(C)$ is relevant here. It is not hard to provide an epistemic rationale for maximising $Pr(P|C)$ in an inference of C

out of P, since it is a central aim of science and philosophy to provide *good* explanations. In the traditional *deductive nomological* model of explanations, the paradigmatic case of a good explanation consists of a deductively valid argument with true laws and auxiliary assumptions as premisses and the claim to be explained as conclusion of the argument (see Hempel 1965). Now, a high likelihood of P given C *approximates* deduction of P from C for which reason maximising $Pr(P|C)$ also serves for approximating the paradigmatic case of a good explanation. But what about $c(C)$? In how far does decreased complexity or increased simplicity serve the epistemic goals of science and philosophy? Clearly, without taking into account $c(C)$ we would lack a criterion of selecting among a multitude of potential explanations. But if it were just for reducing the number of potential explanations then also a random choice would serve the aim. According to the argument above, not considering $c(C)$ would allow for ad hoc explanations. But what is the epistemic rationale of excluding ad hoc explanations? One argument which is brought forward quite often is that ad hoc explanations *overfit* the data and so in case there is some error in the data, ad hoc explanations also fit errors. So, the argument is that since P might contain false values or statements, validating explanations that perfectly explain erroneous P are themselves defective and their explanantia C wrong. Since decreased complexity $c(C)$ allows for avoiding overfitting, less complex Cs are also less prone to fit errors. As the literature on model selection shows, this can provide a rationale for also taking into account $c(C)$ in choosing among accessible potential explanations.

So, abductive inference consists in an inference to the best accessible potential explanation. ‘Best’ is understood as balancing two measures of an explanation of P by help of C: the likelihood $Pr(P|C)$ should be high and the complexity $c(C)$ should be low. An epistemic rationale for the former constraint results from approximating traditional models of explanation. Such a rationale for the second constraint might result from considerations of the literature on model selection showing that $c(C)$ influences C’s proneness of overfitting, and by this C’s proneness of also fitting errors. In the following section we are going to make this argument in favour of minimising $c(C)$ explicit.

7.2 The Epistemic Value of Simplicity

One way of arguing for minimising $c(C)$ is to postulate as aim of science and philosophy not only to provide true explanations, but also non-ad hoc, universal, simple ones. In this way already by convention about the aim of science and philosophy a demand of minimising $c(C)$ follows. However, there is also the possibility of trying to reduce the epistemic value of minimising $c(C)$ to the epistemic value of providing true explanations. The most famous approach in this direction is an application of an information theoretical framework to the problem of *ad hoc* explanations. The main line of argumentation is as follows (see Forster and Sober 1994):

1. Data P might be noisy and involve *error*.
   Schematically: $Error$

2. An accurate fit of an explanans C to the data P fits also error, it overfits the data.
   Schematically: $Error \Rightarrow (Accuracy \Rightarrow Falsehood)$

3. Whereas a less accurate fit of C to P may depart from error: Closeness to the truth is different from closeness to the data.
   Schematically: $Error \Rightarrow (Inaccuracy \Rightarrow PosTruth)$

4. Fact: The more parameters an explanans C has, the more prone it is to overfit P.
   Schematically: $(Complexity \Rightarrow Accuracy)$ & $(Simplicity \Rightarrow Inaccuracy)$

5. Hence: Simplicity in the sense of having less parameters may account for inaccuracy w.r.t. data P in order to achieve accuracy w.r.t. the truth. So, simplicity is instrumental for truth.
   Schematically: $(Complexity \Rightarrow Falsehood)$ & $(Simplicity \Rightarrow PosTruth)$

As the rough schema shows, this argument is valid along general lines. But what about the truth of the premisses? Considering applications of the abductive methodology to the natural sciences, premise 1, the assumption of error in the data, is a very natural assumption. But then also premise 2 and 3 are straightforward: Assuming that P contains errors one only has a chance of achieving the truth by deviating from P. Intuitively and qualitatively speaking, premise 4 is also straightforward: If an explanans is complex, it allows for fitting a simple as well as a complex explanandum. If an explanans is simple, it might fit a simple explanandum, but it cannot fit a complex one. But clearly, this is an argument too coarse-grained in order to be convincingly applied for quantitative considerations regarding minimising $c(C)$. However, there is also a much more fine-grained version of premise 4 stemming from the literature on model selection and curve fitting—here we focus on the latter, since it became a quite influential approach to the epistemic value of simplicity (see Forster and Sober 1994).

For illustrative purposes we will make only very simplified considerations here. The idea of model selection is that, given a data set $X = \{x_1, \ldots, x_n\}$, one is looking for a curve $F = \{f_1, \ldots, f_n, \ldots\}$ that adequately fits X. Now, it is assumed that X might contain errors, so X deviates from the truth $T = \{y_1, \ldots, y_n, \ldots\}$ (see premise 1). Clearly, the perfect choice would be $F = T$, regardless of X, but since only X is available to us, we have to base our choice of F on X. It is also clear that for any data set X with $n = |X|$ elements choosing as F a polynomial of degree $n-1$ allows one to perfectly fit X. One can always find parameters $a_{n-1}, \ldots, a_0$ such that for all $x \in X$ there is a $z \in \mathbb{R}$: $\langle z, x \rangle \in F$, given $F(z) = x = a_{n-1} \cdot z^{n-1} + \cdots + a_1 \cdot z^1 + a_0$. So, n parameters $(a_{n-1}, \ldots, a_0)$ allow for defining an F that perfectly fits X. If F has less parameters than n, then there are cases where one cannot fit F perfectly to X. So, the number of parameters of F determines possibilities of perfect fitting. However, fitting X perfectly might deviate from the truth T, whereas fitting X imperfectly might allow for achieving the truth T (see premisses 2 and 3). Figure 7.2 depicts this possibility.

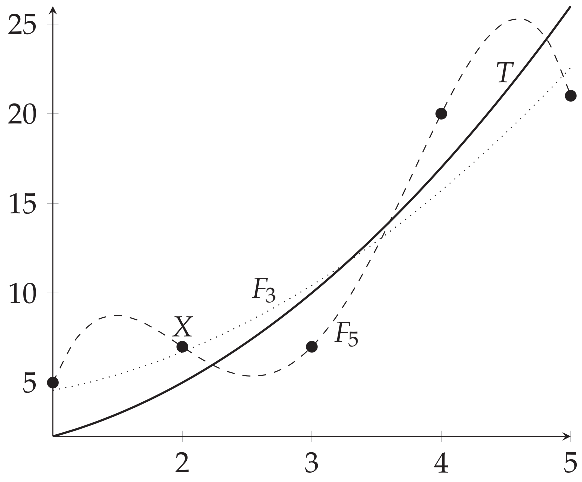

Figure 7.2: Curve fitting with a polynomial of degree 4 with 5 parameters F_5 and a polynomial of degree 2 with 3 parameters F_3. F_5 perfectly fits data set X, whereas F_3 deviates from X. However, F_5 has more distance from the truth T, whereas F_3 approximates T.

Clearly, the advantage of not overfitting by a simpler model (like F_3 in figure 7.2) compared to a more complex model (like F_5 in figure 7.2) depends on our choice of error, namely the distance between X and T. If there were no error ($X \subseteq T$), then the more complex model F_5 would be better off than the simpler one F_3. However, a famous result of Hirotugu Akaike shows that on average (i.e. in estimating) simplicity matters. Forster and Sober (1994) have transformed Akaike's result to the philosophical debate of problems surrounding curve fitting. The result is as follows (see Forster and Sober 1994, p.10): The estimated predictive accuracy of the family of a model F given some data X, which is also called the *Akaike information measure according to the Akaike information criterion* $AIC(F, X)$, is determined by:

$$AIC(F, X) \propto \log(Pr(X|F)) - c(F) \tag{AIC}$$

Where $c(F)$ is the number of parameters of F (i.e. the degree of the polynomial F plus 1) and F is supposed to be most accurately parameterised regarding X (i.e. it is the/a polynomial of degree $c(F)$ that is closest to X in terms of the the sum of squares of the differences).

Note that the idea of the Akaike information criterion is to select an F such that the estimated accuracy regarding the truth T of the family of F given some data X is maximised. Now, as (AIC) tells us, maximising $AIC(F, X)$ is twofold: It consists of maximising the log-likelihood of F given X while at the same time one needs to keep an eye on holding complexity or the number of parameters of F low.

Note also that we have chosen this criterion only as a proxy. In principle any other information criterion, as, e.g., also the Bayes information criterion (BIC) might be employed for cashing out the epistemic value of simplicity (for a discussion of different information criteria see Schurz 2013, sect.5.10.5; and Schurz 2014). What is relevant for our discussion is not which exact criterion one chooses, i.e. which exact balance between $c(C)$ and $Pr(P|C)$ one opts for, but only that there is some balancing going on, i.e. that $c(C)$ is relevant for abduction.

This framework has a wide range of applications. As Forster and Sober (1994) demonstrate, it allows for reducing the value of unification, simplicity of causal models, and the value of non-ad hoc explanations to the epistemic value of gaining truths. Hitchcock and Sober (2004) were able to expand the discussion in such a way that also the problem of *novel facts* can be addressed by help of (AIC): They were able to show that prediction of data is more instrumental for gaining truths than accommodation is. In the following part of this section we are going to briefly present some of these arguments cashing out the value of simplicity in order to provide a gauge for our application to the abductive methodology.

Regarding unification, the problem is as follows: Given two domains/data sets X_1, X_2, one might ask why is it sometimes better to provide a unified model F_X about both domains $X = X_1 \cup X_2$ which might be less accurate than two models F_{X_1} and F_{X_2}, each of which models just the respective domain X_1 and X_2? So, why choose F_X, although $Pr(X|F_X) < Pr(X|F_{X_1} \cup F_{X_2})$? A simple answer in the line of (AIC) is that in general the number of parameters for defining $F_{X_1} \cup F_{X_2}$ will be higher than that of F_X, so the inaccuracy of F_X might be compensated by its simplicity such that $AIC(F_X, X) > AIC(F_{X_1} \cup F_{X_2}, X)$ (see Forster and Sober 1994, sect.3).

In a similar way one can also account for the problem of ad hoc modifications: Poppers critique of such modifications and explanations in terms of his falsificationism were expanded onto a methodological level by Imre Lakatos. Lakatos (1970) suggested to differentiate between innovative and degenerative research programmes, where a research programme consists of a methodology plus a core of axioms T and a periphery of auxiliary and *ceteris paribus* assumptions H. A research programme $\langle\langle T, H_1\rangle, \langle T, H_2\rangle, \dots\rangle$ is called *degenerative*, if modifications of the research programme's periphery from H_n to H_{n+1} tend to decrease its degree of falsifiability only, and *innovative* otherwise. It is in full agreement with Lakatos' proposal to modify the periphery in such a way that the core of the research programme remains untouched. However, if modification of the periphery takes too much effort, i.e. modification in order to avoid falsification becomes the standard response to new (conflicting) data, then it might be better to throw in the sponge and start with a new core establishing a new research programme. Also here (AIC) turns out to be fruitful, inasmuch as it allows for a nice explication of a research programme being degenerative (see Forster and Sober 1994, sect.5): A research programme is *degenerative* iff a loss in simplicity of the programme's core plus periphery is not compensated by a sufficient gain in fit with the data according to (AIC). So, if there is a negative AIC-development in the sense that $AIC(\langle T, H_{n+1}\rangle, X_{n+1}) \;<\; AIC(\langle T, H_n\rangle, X_n)$ for sufficiently many n, then the research programme has to be evaluated as degenerative and a new one has to come in place.

We now want to cash out (AIC) also for reducing the value of simplicity of the abductive methodology presented before—namely the value of $c(C)$ in (Abd)—to the epistemic value of gaining truths. At least at first glance it is quite straightforward to implement (AIC) into the abductive methodology outlined above: The data set X is to be identified with the premise of the abductive inference P, the explanandum. And the conclusion of the abductive inference C, the explanans, is to be identified with the curve that tries to fit X, i.e. F. The result is an Akaike-motivated characterisation of abductive reasoning: Assume that the only available *potential explanantia* are $C_1, \dots, C_n$. Then it holds:

$$\begin{aligned} &C_i \text{ can be inferred from } P \text{ by abduction iff}\\ &\qquad\text{for all } j \in \{1, \dots, n\}:\\ &\log(Pr(P|C_i)) - c(C_i) \;\geq\; \log(Pr(P|C_j)) - c(C_j)\\ &\text{(In case more than one } C_i \text{ satisfy this constraint}\\ &\qquad\text{one might freely choose among them.)} \end{aligned} \tag{AIC-Abd}$$

According to this characterisation, every inference to an explanation is abductively permitted which manages to get the best balance between likelihood and simplicity. Note that since $Pr(P|C) \in [0, 1]$, $\log(Pr(P|C)) \in (-\infty, 0]$. Furthermore, in principle the complexity of C might have no upper limit (F might be a polynomial of arbitrarily high degree), so $c(C) \in [0, +\infty)$. So, in trying to maximise $Pr(P|C)$ and minimise $c(C)$ one also tries to maximise $\log(Pr(P|C)) - c(C)$.

Clearly, (AIC-Abd) also satisfies the constraint (Abd). Furthermore, one can easily think of examples where (Abd) fails to license or exclude an inference, whereas (AIC-Abd) does allow for it. If, e.g., the only potential explanantia are C_1, C_2, then the values provided in table 7.1 do not allow for validating an explanation by (Abd), but by (AIC-Abd). So, (AIC-Abd) is stronger than (Abd):

Expl.	$Pr(P\|C)$	$c(C)$	$\log(Pr(P\|C)) - c(C)$	(Abd)	(AIC-Abd)
C_1	0.75	3	-3.1		✓
C_2	0.85	5	-5.1		×

Table 7.1: Example of the decisiveness of (AIC-Abd) compared to (Abd) with a $\log_{10}$-likelihood

Both (AIC) as well as (AIC-Abd) allow for shifting the balance between the (log-)likelihood and simplicity by help of choosing different log-functions. E.g., $c(C)$ has relative much impact if one chooses $\log_{10}$. Since $\log_{10}(0.1) = 1$, if one compares only Cs with likelihoods $Pr(P|C) \geq 0.1$, then only $c(C)$ matters. On the other hand, if one chooses, e.g., $\log_{1.001}$, then $c(C)$s influence almost vanishes: Since $\log_{1.001}(0.94) = -61.91$ and $\log_{1.001}(0.95) = -51.31$, given a difference of $|Pr(P|C_1) - Pr(P|C_2)| = 0.01$ close to the upper bound, only a difference of $|c(C_1) - c(C_2)| > 10$ allows for some impact of simplicity which is, speaking in terms of degrees of polynomials, huge. This shows that there is some room for parameterising balancing of likelihood and simplicity according to (AIC) as well as (AIC-Abd).

The implementation of (AIC) in agreement with (Abd) in the criterion (AIC-Abd) seems to do the job of reducing the value of simplicity (low $c(C)$) to the epistemic value of gaining truth. However, this reduction hinges on several assumptions: First of all, in the case of curve fitting the setting consisted of values of or functions on $\mathbb{R}$: The truth T was supposed to contain values of or to be a function on $\mathbb{R}$; similarly for the selected curve or model F; and also the data set X was supposed to contain elements of or be a partial function on $\mathbb{R}$. Now, in the case of abductive reasoning, our straightforward characterisation Akaike style above consists of sets of propositions P and C. So, in order to be not only an empty formal analogy, there is a need of also linking P and C with (partial) functions on $\mathbb{R}$ too. Second, such a linkage needs to allow for defining a measure of simplicity or complexity $c(C)$ of an explanans C which can be interpreted as the degree of a polynomial. Finally and most importantly, the rationale of taking care of such a measure $c(C)$ is provided by the assumption of error in the data. So one also needs to provide an interpretation of error in such a setting. We will fill in further details when we come to the meta-inductive justification of abduction in section 7.4. However, before that we want to characterise the second kind of abduction, creative abduction, in detail.

7.3 Creative Abduction

Let us come to the second form of abduction, namely creative abduction. As we have mentioned above, creative abductions can be thought of as creating at least some of the hypotheses, explanations, predictions, or theories which are in a second step assessed by selective abduction. They are intended as inferences for generating hypotheses featuring new theoretical concepts on the basis of empirical data. Now, theoretical concepts are intimately connected to empirical phenomena via dispositions (see, e.g., Carnap 1936, 1937), for which reason creative abduction particularly focuses on approaching empirically correlated dispositions. Schurz (2008a) differentiates between different patterns of abduction and argues for the view that at least one kind of creative abduction can be theoretically justified. In a nutshell, his approach is based on the idea that inferences to theoretical concepts unifying empirical correlations among dispositions can be justified by Reichenbach's (1971) *principle of the common cause*. The details of this approach are spelled out by help of the framework of Bayesian networks in (Feldbacher-Escamilla and Gebharter 2019). This framework, if causally interpreted, can be seen as a generalisation of Reichenbach's ideas (see Glymour, Spirtes, and Scheines 1991). In the following we characterise creative abduction. We will proceed as follows: First, we briefly describe the approach of Schurz (2008a) to creative abduction and how it allows for unifying strict empirical correlations among dispositions. Afterwards, we show how successful cases of creative abduction can be modelled within the more general framework of Bayesian networks.

Following (Schurz 2008a; and Schurz 2016), we focus on a simple analysis of dispositions as introduced by the early logical empiricists (see Carnap 1936, 1937, e.g.). According to this analysis, whether an object x has a disposition D depends on whether certain test conditions T lead to a specific reaction R. For an object x to be soluble in water, for example, it is required that x dissolves at some time t if put into water at t:

$$\forall t(T(x,t) \rightarrow (D(x) \leftrightarrow R(x,t))) \tag{7.1}$$

According to the traditional understanding, T and R are empirical concepts, while the dispositional concept D is a not directly observable theoretical concept. Note that equation (7.1) comes close to a partial definition of D on the basis of T and R, except that the dispositional term is not relativised to t. It is a well-known fact that the only non-conservative (or creative) import of equation (7.1) is the following assumption about the uniformity of test-reaction pairs (see Feldbacher-Escamilla 2020b, sect.3.3): If at some time t an object x satisfies the test conditions and brings about the corresponding reaction, then x will do so at any time t:

$$\exists t(T(x,t)\&R(x,t)) \rightarrow \forall t(T(x,t) \rightarrow R(x,t)) \tag{7.2}$$

Note that equation (7.1) and equation (7.2) are empirically equivalent. If equation (7.2) has been established on empirical grounds, then introducing a disposition D via equation (7.1) is a theoretical means to explain equation (7.2). However, not much is gained by introducing D since for each regularity among test-reaction pairs a distinct disposition has to be postulated. Things become more interesting once we focus on regularities among several dispositions $D_1, \ldots, D_n$, each characterised by a corresponding test-reaction pair consisting of T_i and R_i (with $1 \leq i \leq n$). Now assume that we found strict pairwise empirical correlations between all of these dispositions $D_1, \ldots, D_n$, meaning that

$$D_i(x) \leftrightarrow D_{i+1}(x) \text{ for all } 1 \leq i < n. \tag{7.3}$$

This amounts to the assumption that the following statement has been empirically established:

$$\exists t(T_i(x,t) \& R_i(x,t)) \rightarrow \forall t(T_j(x,t) \rightarrow R_j(x,t)) \text{ for all } 1 \leq i, j \leq n \tag{7.4}$$

Let us call each statement of this form a *crossed uniformity assumption*. Given n test-reaction pairs for n dispositions $D_1, \ldots, D_n$, we get n^2 such crossed uniformity assumptions (Schurz 2008a, p.226). It is a logical fact that this is empirically equivalent to introducing one higher-level dispositional concept $\mathcal{D}$ characterised by n test-reaction pairs:

$$\forall t(T_i(x,t) \rightarrow (\mathcal{D}(x) \leftrightarrow R_i(x,t))) \text{ for all } 1 \leq i \leq n \tag{7.5}$$

Note that introducing the theoretical concept $\mathcal{D}$ via equation (7.5) reduces the number of law statements from n^2 to n. In this sense such a reduction can be understood as unificatory. The abductive inference consists in the introduction of $\mathcal{D}$ via equation (7.5) on the basis of equation (7.4). It can be illustrated on the following example inspired by (Hempel 1965): Assume that at some time the inhabitants of a not too distant possible world realised that some objects have the disposition to attract iron (D_1) and that some objects have the disposition to produce electricity when moved along a wire (D_2), meaning that they introduced the two theoretical concepts D_1 and D_2 on the basis of equation (7.2) and in accordance with equation (7.1). Suppose further that both discoveries were made independently of each other, but that people found out later on that the dispositions D_1 and D_2 are correlated (equation (7.3)) via observing that their corresponding test and reaction conditions coincided (equation (7.4)). They might then have started to explain this correlation by introducing the higher-level disposition of generating an electromagnetic field $\mathcal{D}$ via equation (7.5).

Note that creative abduction as discussed above can be interpreted either in a realist or an instrumentalist way. Under the latter interpretation $\mathcal{D}$ is taken to be nothing over and above a more or less useful theoretical means to unify empirical descriptions of certain phenomena of interest that can—in principle—be replaced by any other concept with equal empirical adequacy and unificatory power. Under the realist interpretation, on the other hand, $\mathcal{D}$ is assumed to represent a real structure; statements involving $\mathcal{D}$ are considered to be either true or false. Schurz (2008a) as well as Schurz (2016) made a strong case in favour of a realist interpretation by endorsing the common cause principle of Reichenbach (1971):

(CCP) If two properties P and Q are correlated and neither P causes Q nor Q causes P, then P and Q are effects of a common cause D.

The common cause principle (CCP) demands that every correlation among any pair of properties not standing in direct causal dependence to each other has to be explained by the existence of an independent common cause. In this sense (CCP) provides a way of causally unifying observed regularities. In the case of pairwise empirically correlated dispositions such as $D_1, \ldots, D_n$ above, (CCP) supports a realist interpretation of the unifying higher-level disposition $\mathcal{D}$: The correlation among dispositions $D_1, \ldots, D_n$ is explained by postulating a common cause $\mathcal{D}$.

Now, this characterisation of creative abduction can be generalised by embedding it into the more general framework of Bayesian networks. Furthermore, the generalisation allows for several interpretations of creative abduction. One might consider common cause abduction as a realist inference method, but one might also consider it as an instrumentalist one. This is, because, though Bayesian nets can be causally interpreted, one does not have to subscribe to a realist interpretation when employing this particular framework to model creative abduction (for an argument supporting a realist interpretation of the causal Bayesian network framework, see Gebharter 2017; Schurz and Gebharter 2016). An instrumentalist can still use Bayesian networks without a causal interpretation as a tool for making abductive inferences featuring unification.

Let us now come to the model of creative abduction in the Bayesian network framework. We represent pairwise empirically correlated lower-level dispositions by variables $D_1, \ldots, D_n$ and the abduced higher-level disposition by a variable $\mathcal{D}$. Evidence for one of the lower-level dispositions D_i (with $1 \leq i \leq n$) is represented by a variable E_i which stands for an inductive generalisation of instances of test-reaction conditions such as $(T_i(a_1, t_1) \& R_i(a_1, t_1)) \& \cdots \& (T_i(a_k, t_l) \& R_i(a_k, t_l))$. The dependence of each lower-level disposition D_i on its corresponding evidence E_i is represented the same way as the dependence of a hypothesis on its evidence is typically modelled in the Bayesian framework: For each pair D_i, E_i we draw an arrow $D_i \longrightarrow E_i$. Since the creative abductive step is conducted by applying (CCP) in the approach of Schurz (2008a), we introduce the higher-level disposition variable $\mathcal{D}$ as a common parent of the lower-level disposition variables $D_1, \ldots, D_n$. The resulting graph is depicted in figure 7.3.

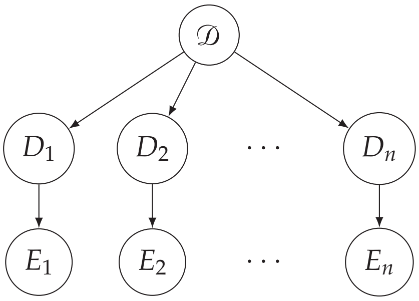

Figure 7.3: Bayesian network for modelling successful instances of creative abduction

Probability flow between dispositions $D_1, \ldots, D_n$ is established via $\mathcal{D}$ if the following general conditions are satisfied:

1. $\mathcal{D}$ is not extreme, i.e., $0 < Pr(\mathcal{D}) < 1$.
2. Each D_i depends positively on $\mathcal{D}$, i.e., $Pr(D_i|\mathcal{D}) > Pr(D_i)$.

From 1 and 2 it follows that $Pr(D_i|D_j) > Pr(D_i)$ if $i \neq j$ (for a proof see, e.g., Dardashti et al. 2019). To account for the corresponding correlations between the evidence $E_1, \ldots, E_n$, the following condition has to be satisfied as well:

3. Each E_i depends positively on its corresponding D_i, i.e., $Pr(E_i|D_i) > Pr(E_i)$.

From 1, 2, and 3 it follows that $Pr(E_i|E_j) > Pr(E_i)$ if $i \neq j$.

Conditions 1, 2, and 3 are necessary conditions for successful creative abduction: They guarantee pairwise correlations among lower-level dispositions that have to be inductively inferred on the basis of observed evidence and build the basis for introducing the higher-level disposition $\mathcal{D}$ which is then, in turn, used to explain these correlations.

Like in the approach of Schurz (2008a), creative abduction provides unification if modelled Bayesian style. In the original (deterministic) approach introducing the higher-level disposition $\mathcal{D}$ provided unification of n^2 empirical law statements establishing pairwise empirical correlations among n lower-level dispositions to n higher-level dispositional statements. In the Bayesian setting, pairwise empirical correlations between n lower-level dispositions $D_1, \ldots, D_n$ consist in $\binom{n}{2}$ probabilistic dependencies of the form $Pr(D_i|D_j) > Pr(D_i)$, where $1 \leq i \neq j \leq n$. Similarly, for the dependencies among pairs of evidential variables there are $\binom{n}{2}$ empirical correlation statements of the form

$$Pr(E_i|E_j) > Pr(E_i), \text{ where } 1 \leq i \neq j \leq n. \tag{7.6}$$

It follows from the Markov factorisation (equation (2.1)) that these $\binom{n}{2}$ empirical correlation statements can be unified by the $2n + 1$ probabilistic statements in conditions 1, 2 and 3.: n statements of the form $Pr(E_i|D_i) > Pr(E_i)$ (with $1 \leq i \leq n$), n statements of the form $Pr(D_i|\mathcal{D}) > Pr(D_i)$ (with $1 \leq i \leq n$), and 1 statement $0 < Pr(\mathcal{D}) < 1$. To compare the approach of Schurz (2008a) and the Bayesian approach w.r.t. their unificatory power, we introduce a simple measure u intended to capture the intuitions about unification outlined above. Given n correlated lower-level dispositions, $u(n)$ measures the ratio between $x(n)$ empirical statements to be unified and $y(n)$ unifying theoretical statements. In order to shift the neutral case to 0, we subtract from this ratio 1:

$$u(n) = \frac{x(n)}{y(n)} - 1$$

Its output is in the interval $[-1, \infty)$, where a negative value means that the theoretical description is more costly than simply listing the empirical statements, 0 means that there is no gain but also no cost in providing a theoretical description, and a positive value means that the theoretical description provides unification. This kind of measuring unificatory power by counting statements, argument patterns, etc. is common in the unification literature (see Woodward 2018, sect.5.4). There are, however, also other ways of measuring unificatory power. To avoid problems Bayesian measures have with common cause structures (see Schupbach 2005), Myrvold (2017) suggests to avoid an explicit representation of common causes. For purposes of unification, one should use hypotheses postulating such common causes instead. But since we focus on creative abduction here, avoiding common causes in order to maintain a Bayesian measure for unification seems to be inappropriate for our endeavour. For this reason and in order to compare the Bayesian network analysis with Schurz (2008a), we decided in favour of a simple counting measure.

A comparison of the unificatory power of both, the original and the Bayesian network approach, is provided in figure 7.4 (thin and thick solid line): In the case of strict (unconditional) correlations, the original approach fares better than the Bayesian approach. This is due to the theoretical power of the Bayesian framework which requires more parameterisation.

Up to now we focused on comparing unification of statements about *unconditional* empirical correlations. However, much more empirical correlations are possible in the Bayesian setting. If the evidential base is strictly correlated (i.e., $Pr(E_i|E_j)$ and $Pr(E_i|\overline{E}_j)$ with $1 \leq i, j \leq n$ are extreme), then it follows from the Markov factorisation (equation (2.1)) and conditions 1, 2, and 3 that each two variables E_i, E_j (with $i \neq j$) are independent conditional on any set of other evidential variables. Thus, the unconditional dependence statements in equation (7.6) capture all dependencies among variables $E_1, \ldots, E_n$ in this setting. However, if some correlations among pieces of evidence cannot be screened off by some non-empty set of other evidential variables, then also many *conditional* empirical dependencies may hold among pairs of evidential variables. In particular, there can be up to $2^{n-2} \cdot \binom{n}{2}$ empirical dependencies of the form

$$\begin{aligned} Pr(E_i|E_j, \mathbf{Z}) > Pr(E_i|\mathbf{Z}), \text{ where} \\ 1 \leq i \neq j \leq n \text{ and } \mathbf{Z} \subseteq \{E_k : 1 \leq i \neq k \neq j \leq n\}. \end{aligned} \tag{7.7}$$

If these conditional dependencies are also taken into account, then creative abduction Bayesian style provides a tremendous gain in unificatory power (see figure 7.4, thin and thick solid as well as the dashed line). From 1, 2, and 3 it also follows that $Pr(E_i|\mathbf{Y}) > Pr(E_i|\mathbf{Z})$, where $\mathbf{Z} \subset \mathbf{Y}$ and $\mathbf{Y}$ are sets of evidential variables different from E_i (for a proof see, e.g., Dardashti et al. 2019). So, the Bayesian network framework allows for a much more fine-grained modelling of non-strictly empirically correlated dispositions which can be found in many higher-level sciences such as economics, medicine, psychology, and sociology.

In the preceding two sections we have characterised selective abduction. We have already provided a partial justification of selective abduction by highlighting the epistemic value of simplicity. In this section we have characterised creative abduction. We have also spelled out wherein simplicity of creative abduction consists in, namely in unification as measured via u. Unification in this sense is expressed by the ratio of the number of statements to be explained (*explanandum*/*explananda*) and the number of statements used for an explanation (*explanans*/*explanantia*). This is the way we measured simplicity in the case of creative abduction. However, note that there is a difference between simplicity in the case of creative abduction (u), and epistemically justified simplicity in the case of selective abduction (complexity c which is the degree of the polynomial for curve fitting). In the next section we will show how simplicity in the case of creative abduction in the Bayesian framework translates into simplicity as used in the case of selective abduction. Afterwards, we complete the justification of these abductive methods by help of meta-induction.

7.4 A Meta-Inductive Justification of Abduction

In this section we complete the characterisation of selective and creative abduction and then apply the theory of meta-induction in order to show how they can be justified. For this purpose we need to fulfil two tasks: First, we need to explain how unification provided by creative abduction translates into simplicity as discussed in the information theoretical approach Akaike style, and hence is epistemically relevant. And second, we need to show how abductive selection of creative abduction amongst others needs to be

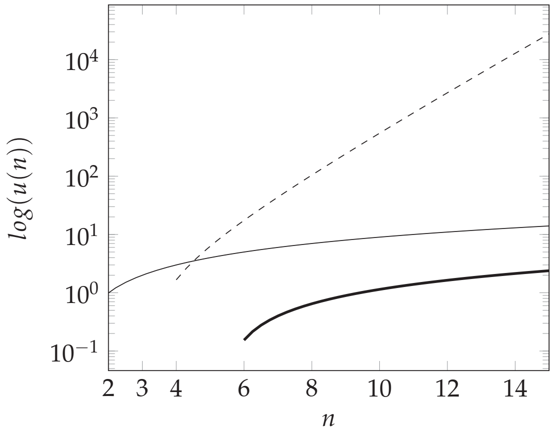

Figure 7.4: Comparison of unificatory power in the original and in the Bayesian setting: n is the number of pairwise empirically correlated dispositions. $u(n)$ measures the unificatory power given n such dispositions by taking the ratio between the number of their corresponding empirical law statements and the number of unifying statements with a shift of the neutral case to 0. In the original setting (thin solid line), $u(n)$ is calculated via $\frac{n^2}{n} - 1$, where n^2 is the number of empirical law statements in equation (7.4). The unifying statements consist of the n formulae in equation (7.5). In the Bayesian setting (thick solid line), the corresponding $u(n)$ is calculated via $\frac{\binom{n}{2}}{2n+1} - 1$. The nominator $\binom{n}{2}$ expresses the number of statements describing the strict (unconditional) empirical correlations in equation (7.6), and the denominator $2n + 1$ is the number of unifying statements in conditions 1, 2, and 3. The unificatory power $u(n)$ in the Bayesian setting with conditional dependencies (dashed line) is calculated via $\frac{2^{n-2} \cdot \binom{n}{2}}{2n+1} - 1$. The numerator expresses the number of statements describing the conditional and unconditional dependencies according to equation (7.7), and the denominator $2n + 1$ is, again, the number of unifying statements in conditions 1, 2, and 3. This case shows that once one allows for non-strict (conditional) correlations, then abductive inferences in the Bayesian network setting receives a tremendous boost in terms of unificatory power—note that the y-axis plots the logarithm of the ratio with a shift of the neutral case to 0.

done in detail in order to guarantee long run optimality which suffices for epistemic justification.

Let us begin with the first problem: How does unification u of the Bayesian framework translates into simplicity (inversely proportional to c)? Forster and Sober (1994) discuss already the question why, given an effect E, explanations that postulate fewer causes should be preferred over explanations that postulate more. If we assume two binary causal variables ($val(D_1), val(D_2) \in \{0, 1\}$) then the cases to be considered are as follows:

$Pr(E\|\cdot)$	D_1	$\overline{D_1}$
D_2	$d_0, d_1, d_2, d_{1,2}$	d_0, d_2
$\overline{D_2}$	d_0, d_1	d_0

Where the d_is are the parameters of the models. Now, given these ingredients, one can formulate several models as, e.g.:

(CE1) $c = 1$: $Pr(E|D_1, D_2) = d_0 + d_1 \cdot val(D_1)$

(CE2) $c = 2$: $Pr(E|D_1, D_2) = d_0 + d_1 \cdot val(D_1) + d_2 \cdot val(D_2)$

(CE3) $c = 3$:
$Pr(E|D_1, D_2) = d_0 + d_1 \cdot val(D_1) + d_2 \cdot val(D_2) + d_{1,2} \cdot val(D_1) \cdot val(D_2)$

Here (CE1) states that only a single cause (D_1) is relevant for the explanation of E. (CE2) states that two causes (D_1, D_2) are relevant, but that these do not interact. And (CE3) states that the same two causes are relevant, but that they also interact. By similar reasoning as in section 7.2 where we discussed the Akaike framework, it might be the case that the assumption of interactive causes (CE3) provides a better explanation in terms of accuracy. However, since the number of parameters is also increasing from (CE1) to (CE3), better fit of the models increases also their proneness for overfitting, i.e. fitting errors in E. If one assumes in particular that the accuracy of these three models is equal, then (AIC) tells us to favour the simpler model, since then: $AIC(CE3, E) < AIC(CE2, E) < AIC(CE1, E)$ (see Forster and Sober 1994, sect.4).

Now, we *cannot* translate the measure for unification u of the preceding section directly into a measure of the number of parameters needed for such models c. However, we can provide a transformation which preserves, considering the relevant cases only, the comparative structure. This means that for the relevant cases we cannot provide a transformation of u to c on the cardinal scale, but we can provide such a transformation on the ordinal scale. Here is an outline of how it works: Again let $\binom{n}{2}$ be the number of statements about strict and unconditional empirical correlations according to equation (7.6). Then a Bayesian network as depicted in figure 7.3 allows for explaining these correlations by help of $2n + 1$ statements, namely one statement for the non-extremity of the common cause $\mathcal{D}$ (condition 1), n statements for the positive correlation of the intermediate dispositions D_i with respect to the higher-level disposition $\mathcal{D}$ (condition 2), and n statements for the positive correlation between the evidence E_i with the respective D_i (condition 3). So, this explanation has unificatory power $u(n) = \binom{n}{2}/(2n + 1) - 1$.

In the Bayesian setting, one can account for explaining the empirical correlations by help of many other networks. So, e.g., one could account also for a pairwise correlation among the Es by postulating common causes for pairs of Es which in turn are again linked in pairs by common causes and so forth until all are linked via a root cause $\mathcal{D}$. Figure 7.5 illustrates such a Bayesian network. Now, there are $n - 1$ variables $D_1, \ldots, D_{n-1}$ (where $D_{n-1} = \mathcal{D}$) in such a network one considers as a common cause (for combinatorial considerations on this see our discussion of binary decision trees in section 3.3). In order to account for empirical correlations in such a network, one would need to postulate conditions similar to 1–3 for the common causes, namely first, non-extremity of $\mathcal{D}$: $0 < Pr(\mathcal{D} < 1$. And second, pairwise positive correlations between children and parents: $Pr(E_1|D_1) > Pr(E_1)$, $Pr(E_2|D_1) > Pr(E_2)$, $\ldots$ as well as $Pr(D_1|D_{\frac{n}{2}+1}) > Pr(D_1)$, $Pr(D_2|D_{\frac{n}{2}+1}) > Pr(D_2)$, $\ldots$. As can be seen according to figure 7.5, the latter amounts to stating such a correlation for any arrow. Since there are $n-1$ Ds and each D has two arrows, there are $2 \cdot (n-1)$ such arrows. So, taking these correlation statements plus the one statement for non-extremity of $\mathcal{D}$, one needs $2n-1$ explanantia statements. Hence, for such an abductive inference $u(n) = \binom{n}{2}/(2n-1) - 1$.

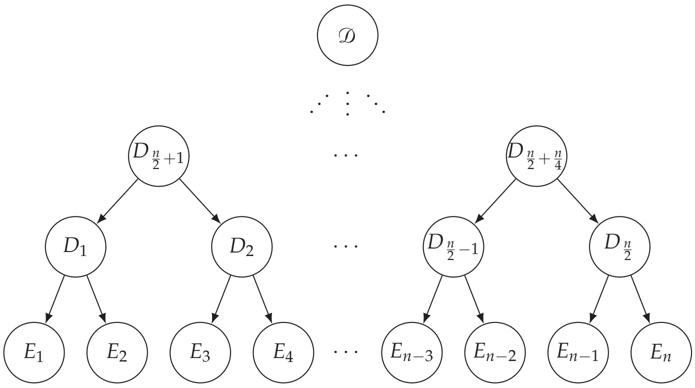

Figure 7.5: Bayesian network for modelling creative abduction with complex theoretical structure

It is clear that the one can increase the performance of the Bayesian approach to abduction by cutting intermediary theoretical terms little by little. One can optimise this approach by omitting all the intermediate lower-level dispositions (the Ds) and unify the correlations among the evidence $E_1, \ldots, E_n$ by directly abducing $\mathcal{D}$. The corresponding Bayesian network's graph is depicted in figure 7.6. In order to account for the empirical correlations in such a network, one needs to assume condition 1 ($0 < Pr(\mathcal{D}) < 1$) and n times a positive correlations between the Es and $\mathcal{D}$: $Pr(E_1|\mathcal{D}) > Pr(E_1), \ldots, Pr(E_n|\mathcal{D}) > Pr(E_n)$. Hence, the unificatory power is: $u(n) = \binom{n}{2}/(n+1) - 1$.

In general, introducing intermediary lower-level theoretical terms (dispositions) comes at the expense of unificatory power. However such dispositions $D_1, \ldots, D_n$ are sometimes practically necessary to find a more general higher-level disposition $\mathcal{D}$. So, sometimes for practical reasons the unificatory payoff is diminished. To illustrate this by help of an example, one might think of the history of electromagnetism (this is a very rough sketch backing on Verschuur 1993): Already in ancient times attraction of

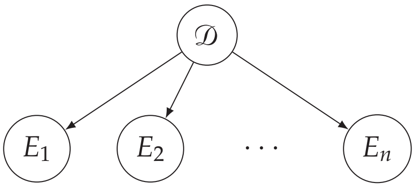

Figure 7.6: Bayesian network for modelling creative abduction with the most simple theoretical structure

iron by lodestone E_1 was correlated with different behaviour E_2 of north and south pole of such stones (to attract and to repel each other) via a disposition *magnetism* D_1. In medieval times, geographic directions E_3 were correlated with the displaying of compass needles (E_4) via a disposition *compass* D_2. In the modern era these two dispositions were correlated via a higher order disposition *magnete tellure* D_3 (by William Gilbert who speaks of "the great magnet earth"). In the nineteenth century this disposition was correlated with dispositions in the realm of electricity to a higher order disposition D_4 (by Hans Christian Ørsted, André-Marie Ampère, and Michael Faraday), which was in turn correlated with further dispositions of optics by James Clerk Maxwell via the higher level disposition of *electromagnetism* D_5. Correlating via higher level dispositions still goes on and on (*quantum electrodynamics* etc.). All these theories achieve better and better unification and allow for cutting out more and more intermediary links (e.g., magnetism, electricity, and optics are treated within one framework of electromagnetism). Figure 7.7 illustrates these different performances: Abducing one common cause which allows for explaining all the empirical correlations as in the Bayesian network of figure 7.6 fares best (dotted line). Abducing common causes in a pairwise manner as in the Bayesian network of figure 7.5 fares better (dashed line). And finally, abducing a lower level disposition for each empirical phenomenon and then unify them by help of a higher level disposition also allows for unification, but fares worst (thick line—this is the same abduction as the thick line in figure 7.4).

Now, as we can see from the *structural equations* (CE1)–(CE3), with an increasing number of Ds, also the number of parameters in the polynomial increases. Note that for n empirical variables, the number of theoretical variables (disposition variables) in figure 7.3 is $n + 1$. The number of theoretical variables in the complex Bayesian network of figure 7.5 is $n - 1$. The number of theoretical variables in the simplest Bayesian network of figure 7.6 is 1. And the number of theoretical variables in any Bayesian network in between which accounts for the empirical correlations and cuts out theoretical variables from the complex network is also in between i.e. $< n - 1$ and > 1. Hence, the degrees of the polynomial in the respective structural equations have the same order. So, simplicity in terms of number of statements in the explanans (u) matches ordinally simplicity in terms of

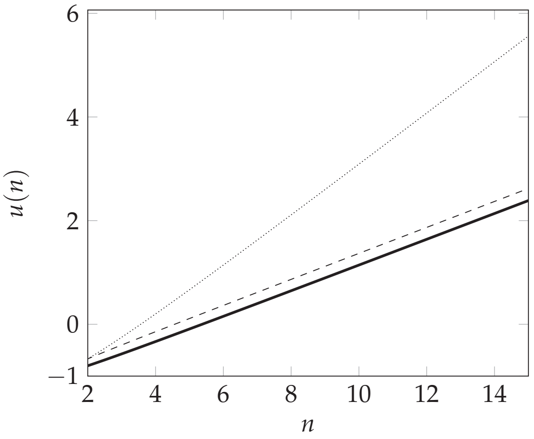

Figure 7.7: Comparison of unificatory power in the Bayesian setting: As in figure 7.4, n is the number of pairwise empirically correlated dispositions and $u(n)$ measures the unificatory power given n. The dotted line represents unification by abducing a single common cause ($u(n) = \frac{\binom{n}{2}}{n+1} - 1$). The dashed line represents unification by abducing a pairwise common cause structure ($u(n) = \frac{\binom{n}{2}}{2n-1} - 1$). The thick line represents unification by abducing first lower level dispositions which are then unified by one higher level disposition ($u(n) = \frac{\binom{n}{2}}{2n+1} - 1$).

the degree of a polynomial (c). Hence, we can also apply the information criterion equation (AIC-Abd) for selecting creative abduction.

So much for the first task, the translation of simplicity in terms of unification u to simplicity in terms of the degree of a polynomial c. This allows for justifying the epistemic relevance of (positive) u. Now, given the epistemic relevance of simplicity, how should we select among hypotheses, explanations, theories? According to selective abduction equation (AIC-Abd), we should try to maximise the information theoretical balance between accuracy ($Pr(P|C)$) and simplicity ($c(C)$). By choosing that hypothesis, explanation or theory which has the *best* balance, we will be closest to the truth, which might be different from being closest to the data P (see Forster and Sober 1994, p.6). So, given the epistemic aim of *being close to the truth*, equation (AIC-Abd) seems to be an optimal means to achieve this end. However, this is with respect to explanation. What about predictions? What about choosing the *best* balanced hypothesis or theory for prediction?

Now, the theory of meta-induction can be applied for optimising predictions in any respect, as long as the formal conditions of the framework are satisfied. In our application to Hume's problem of induction as well as in the motivation of our notions of *regret*, *success*, *learnability*, *optimality*, etc. we interpreted the framework plainly epistemically: Given a prediction game G with $\mathcal{Y}$ and $\mathcal{F}$, we interpreted $\mathcal{Y}$ as the truth and $\mathcal{F}$ as prediction methods or hypotheses about the truth. However, we can also take in a more pragmatic standpoint and interpret $\mathcal{Y}$ as past, present, and future data, and $\mathcal{F}$ as prediction methods or hypotheses about which data will be gathered in the future. Since data typically contains error and noise, it easily falls apart from the truth, hence this interpretation does not coincide with the former. And in this sense it seems to be perfectly fine that also the criteria for success fall apart: Epistemically speaking, we still aim at predictions that are as closest to the truth as possible. However, given our noisy data, we know that we need to aim at predictions that are best balanced between accuracy (fitting) and complexity (overfitting). Success consists not in minimising the distance from the data, but making a prediction which is *best* balanced between these two parameters. So, in order to achieve this goal in the long run, the idea is to use a normalisation of equation (AIC). If r is the highest polynomial we are going to consider and Pr is ϵ-regular (i.e. only $Pr(\bot) = 0$ and all other probabilities are $> \epsilon > 0$—for details see chapter 11), then $AIC(C, P) \in [\log(\epsilon) - r, -r]$. Hence, we can normalise $AIC(C, P)$ to $[0, 1]$ by taking

$$\frac{AIC(C,P) - (\log(\epsilon) - r)}{-\log(\epsilon)}$$

which is in $[0, 1]$.

Now, let us consider a prediction game G with $\mathcal{Y}$, $\mathcal{F}$. Let G be about predicting the *best* balancing for making explanations or predictions. Assume that $\mathcal{Y}$ is a series of objectively *best* balancing. This will still fall apart from providing a most accurate prediction, i.e. a true prediction, because the degree of an extension of a polynomial predicting up to round $t - 1$ might be increased by 1 if one predicts for round t the true value with probability 1, whereas it might not be increased at all by predicting the true value with close to 1 probability and hence deviation from the truth might have a higher AIC (the AIC is higher, if the deviation allows for no change in the degree of the polynomial and the basis of log in equation (AIC) is $> 1/(1 - Pr(P|C))$). Given such an "objective" best balancing, we can interpret $\mathcal{F}$ as a set of theories or hypotheses which provide predictions of what will be the best balance once new data enters the game (i.e. once one moves forward to the next round). We take the predictions in $\mathcal{F}$ to be the actual AICs of the same methods in predicting some event in another prediction game, let us say G'. So, if for $f'_i \in \mathcal{F}'$, $AIC(f'_i, \{Y_1, \ldots, Y_{t-1}\}) = a$, then the respective $f_i \in \mathcal{F}$ predicts for round t as best balance the normalisation of a, i.e. $\frac{a-(\log(\epsilon)-r)}{-\log(\epsilon)}$. G is, so to say, a meta game where any prediction method of the ordinary game G' predicts that it has the right balance for future predictions. In other words, playing G' and making predictions comes with the commitment of claiming also that one's prediction is right in the sense of best balanced—that is a claim in G.

Now, again by success-based mixing of the forecasts about the best balance to be expected, a meta-inductive learner achieves long run optimality in predicting the best balance in G. Now, if we assume that in science creative abductive methods with high unificatory power had the best balance in the past, then using such creative abduction for inferring theoretical frameworks is epistemically justified, since using them is, at the current state of science, the best thing to do: Following the meta-inductive selection allows for optimality in predicting the best balance in G (and actually having the best balance in one's events predictions in G').

Note that given this assumption, anti-abduction fails to be justified: Disunification and theoretically laden hypothesis invention fared suboptimal in past (in G') and hence its predictions of the best balance in G were also wrong. Hence, meta-inductive selection ignores these methods and this is the best thing to do, at least given their past performance. Given this assumption, anti-abduction is by far no optimal means to achieve the epistemic end of being best balanced in G.

A further note is in place: In the argument above we made implicitly the assumption that success in the meta game G and success in G' are synchronous: Whenever one was relatively successful in choosing the right balance for theory and hypothesis invention (G), one also was relatively successful in predicting events (G'). The problem with this assumption is that in principle nothing hinders an adversary in letting fall things apart from each other, and allowing for good performance in G', although failing in G. However, we can argue for our assumption by assuming a past correlation and employing induction (as was justified by meta-induction); by this we can inductively transfer this correlation and are epistemically justified in doing so.

Note that such an approach can be considered as introducing cognitive costs in prediction games. Such an expansion is carried out also, e.g., in (Schurz 2019, sect.7.6).

To briefly sum up: In this chapter we have provided exact characterisations of selective and creative abduction which aim at inferring hypotheses, explanations or theories on the basis of data; the two main relevant factors in doing so are likelihood of the data given the inferred hypotheses and simplicity or unificatory power of the hypotheses; we have provided an argument for the epistemic value of simplicity and unificatory power and have shown how inferences based on them regarding *explanations* allow for optimality justification. Finally, we have also outlined how inferences based on them regarding *predictions* can be justified by employing the framework of meta-induction not only for the likelihood, but also the simplicity factor.

Chapter 8

A Note on Deduction

For further illustration of the meta-inductive justification of induction, this chapter states the problem of justifying deduction analogously to the problem of justifying induction. Afterwards, problems related with a deductive justification of deduction and such problems related with an inductive justification are discussed. Finally, the different epistemic ends which underlie the justifications of different inferences are listed.

Considering the binary case, deductive inferences are characterised by the property of truth preservation with *certainty*; they transfer the truth from the premisses to the conclusion with certainty. Putting forward truth preservation as *the* paradigmatic epistemic end, it seems that deductive inferences are by definition an optimal means to achieve this end. However, in showing this, one presupposes deduction, and so the question is whether a similar problem as in the case of justifying induction shows up also in the case of justifying deduction.

In this chapter we are going to briefly consider the problem of justifying deduction. Since meta-induction and its justification presupposes deduction and is designed for a purpose different than that of truth preservation, we do not think that the theory of meta-induction allows for many insights regarding the justification of deduction. Rather, we think that in restating some problems of justifying deduction in a similar vein as the problem of justifying induction, we can shed further light on the theory of meta-induction.

We will do so by first stating the problem of justifying deduction analogously to Hume's dilemma regarding induction (section 8.1). Afterwards, we briefly discuss problems related to the justification of deduction and link them to the respective problems of justifying induction (sections 8.2 to 8.3). Finally, we briefly compare the different epistemic ends underlying the justification of the epistemic means deduction, induction, and abduction (section 8.4).

8.1 Haack's Dilemma for Deduction

Epistemically speaking, we aim at true (and informative etc.) propositions. Inferences allow us to move from premisses to conclusions, so, if we put forward the ends (true propositions) for looking out for suitable means (inferences) in a strict sense, then we would need to justify inferences via an ability to certainly end up with true conclusions, regardless of the truth of the premisses. Now, clearly "truth generation with certainty" is a task too demanding or restricting in order to be satisfied, for which reason we aim at the next epistemic end, truth preservation with certainty. As stated before, by definition an inference is deductive, if it is truth preserving with certainty. So, it seems that epistemic justification J of deductive inferences can be "readily provided and follows automatically" by definition: By definition deductive inferences preserve the truth from the premisses to their conclusions with certainty. If an inference preserves truth with certainty, then it is justified. Hence, deductive inferences are justified. However, clearly this justification used deductive reasoning (e.g. *modus ponens*), and hence it presupposes what it intended to show in the first place. The problem is, that there seems to be no viable alternative other than using deductive reasoning in justifying deductive reasoning. Now, if we think of induction in the wide sense as the residue class of inferences that are not deductive, then one can put forward the same dilemma we had already in justifying induction also for deduction:

> "Hume presented us with a dilemma: we cannot justify induction deductively, because to do so would be to show that *whenever* the premisses of an inductive argument are true, the conclusion must be true too—which would be *too strong*; and we cannot justify induction inductively, either, because such a 'justification' would be *circular*. I propose another dilemma: we cannot justify deduction inductively, because to do so would be, at best, to show that *usually*, when the premisses of a deductive argument are true, the conclusion is true too—which would be *too weak*; and we cannot justify deduction deductively, either, because such a justification would be *circular*." (see Haack 1976, p.112)

An analogous framing of the problem is explicitly discussed also in (see Jacquette 2011, p.6). It seems that we face the same problem with justifying deduction as we faced already with induction. However, recall our discussion of an inductive justification of induction in section 5.2: There we saw that an inductive justification of induction need not be necessarily circular, if one allows for infinite chains of reasoning and justification. The problem there was that such an infinite reasoning chain can be provided not only for induction, but also for anti-induction. Now, the question is whether the same problem shows up also in a deductive justification of deduction. In the next section we briefly explore this question.

8.2 On a Deductive Justification of Deduction

Consider the following (internal) dialogue: "An inference via *modus ponens* $p, p \to q \vdash q$ is epistemically justified. Why? *Because* it is truth preserving and if an inference as, e.g. modus ponens, is truth preserving, then it is also justified. Why is it truth preserving? Because of its definition (on the basis of a definition of $\to$ e.g. via truth tables). And why is it justified, if it is truth preserving, and if it is truth preserving, then it is justified? *Because* it is truth preserving to infer that it is justified given that it is truth preserving, and if it is truth preserving, then it is justified." Now, one might complain that in justifying modus ponens this way we used modus ponens, and hence the justification is circular. However, we could also say that after the first use of '*because*' we were arguing on a meta level and used another inference rule, structurally equivalent to modus ponens, but not modus ponens itself. Likewise, after the second '*because*', we argued on a meta meta level and used another inference rule, again, structurally equivalent to but not identical with modus ponens and the rule used on the meta level. Since in principle it seems that one could go on with this kind of reasoning ad infinitum, in principle we never need to use one and the same inference rule again, and hence we can avoid a circle. This justification schema is structurally equivalent to that of an infinitist inductive justification as discussed in section 5.2. We have depicted the schema of an infinitist circle-free deductive justification of deduction in figure 8.1.

Here is an explicit level 2 argument for modus ponens of level 1 (here we abstain from considerations regarding truth preservation *with certainty*): '$T(p)$' stands for 'p is true', '$TP(\vdash)$' stands for '$\vdash$ is truth preserving', and '$J(\vdash)$' stands for '$\vdash$ is epistemically justified'.

1. Modus ponens is the rule: $p,\ p \to q \vdash_1 q$ (partial definition of $\vdash_1$)
2. $T(p)$ and $T(p \to q)$. (assumption)
3. If $T(p)$ and $T(p \to q)$, then $T(q)$. (from the definition of $\to$)
4. Hence: $T(q)$. (by modus ponens of level 2 and 2,3)
5. Hence: $TP(\vdash_1)$. (from 1–4)
6. Hence: $J(\vdash_1)$. (from 5 and optimality justification)

This inference on level 2 contains a level 2 modus ponens (step 4) which can be argued for by iteratively adding arguments which are schematically

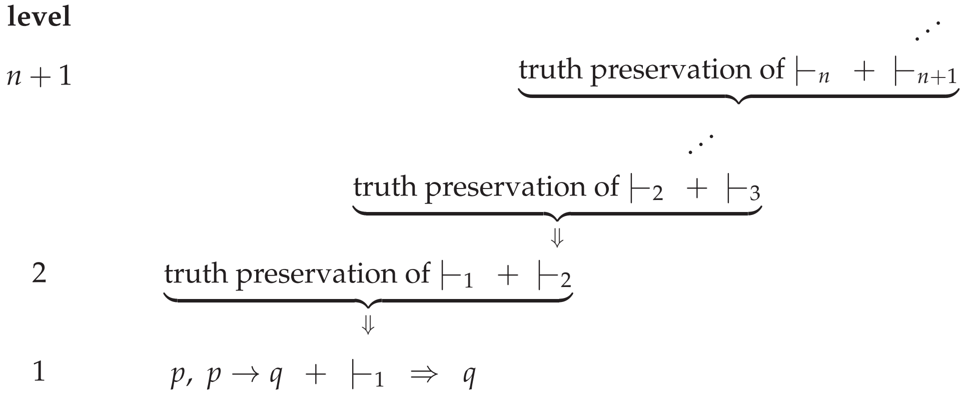
level
$n+1$
truth preservation of $\vdash_n$ + $\vdash_{n+1}$
truth preservation of $\vdash_2$ + $\vdash_3$
2
truth preservation of $\vdash_1$ + $\vdash_2$
1
$p,\ p \to q$ + $\vdash_1$ $\Rightarrow$ q

Figure 8.1: Schema of an infinitist deductive justification of deduction by arguing with deductive rules on different levels: On level 1, e.g., q is inferred from $p, p \to q$ by help of level 1 *modus ponens* which is a rule of $\vdash_1$. On level 2, the truth preservation property of level 1 inferences is used for justifying $\vdash_1$ by help of reasoning with rules of $\vdash_2$. And so on ad infinitum.

equivalent. But note that also many further inferences are used (definitions, conditional proof, optimisation), which opens up a bulk of parallel justification hierarchies needed in order to justify these too. However, this should be not our concern here. We assume that such a parallel justification hierarchy can be established. Rather, we are concerned with the question whether such an infinite hierarchy is free of the problem we saw in the case of such a hierarchy for justifying induction, namely that there is also such a hierarchy for justifying anti-induction.

Now, first note that this kind of justification might seem a little bit artificial, but in fact it has already a long history: So, e.g., we think that Aristotle's concern of vindicating the *principle of non-contradiction* against deductive scepticism is in the line of this argumentation. We do not want to claim that Aristotle actually considered an infinitist notion of *justification* as reasonable—on the contrary, as outlined in chapter 1 he argued quite explicitly against circular reasoning and reasoning by help of infinite regress. However, it seems that in his vindication of the principle of non-contradiction one of his aims was to identify it as a principle whose truth cannot even be questioned without presupposing that it is true. 'Presupposing' is key in ascending from one level to the other, and one might interpret Aristotle as claiming that a deductive sceptic contradicts herself in arguing for the falsity or indeterminacy of this (level 1) principle, since in her arguing she uses the principle (on level 2). Here is a passage which seems to licence such a reading—emphasis by us:

> "It is impossible for anyone to believe that the same thing is and is not, as some consider Heraclitus said—for it is not nec-

> essary that the things one says one should also believe. But if it is not possible for contraries to hold good of the same thing simultaneously (*given that the customary specifications are added to this proposition too*), and the opinion contrary to an opinion is that of the contradictory, then obviously it is impossible for the same person to believe simultaneously that the same thing is and is not; for anyone who made that error would be holding contrary opinions simultaneously. That is why all those who demonstrate go back to this opinion in the end: it is, in the nature of things, the principle of all the other axioms also." (Aristotle 1993, 1005^b22-35)

We need not read Aristotle in the sense of him accepting a direct justification of deduction in an infinitist manner. Rather, it suffices to read him as refuting any critique on the principle of non-contradiction by at least implicitly making the distinction of reasoning levels. This distinction is crucial for the above justification—and in this, and only in this sense, we consider Aristotle in the tradition of such reasoning.

Not only Aristotle, but most famously also Immanuel Kant was concerned with the presuppositions and preconditions for reasoning. In Kantian terms, he was after "*the a priori grounds for the possibility of* ..." ... of everything, amongst others also deduction. Arguments aiming at such a conclusion are so-called *transcendental arguments*, and Kant thought that by such transcendental inferences we end up with a basis which cannot even be questioned without presupposing the basis:

> "The boundaries of logic, however, are determined quite precisely by the fact that logic is the science that exhaustively presents and strictly proves nothing but the formal rules of all thinking [...] in logic, therefore, the understanding has to do with nothing further than itself and its own form." (Kant 1787/1998, BIX, pp.106f)
>
> "Concerning the mere form of cognition (setting aside all content), it is equally clear that a logic, so far as it expounds the general and necessary rules of understanding, must present criteria of truth in these very rules. For that which contradicts these is false, since the understanding thereby contradicts its general rules of thinking and thus contradicts itself." (Kant 1787/1998, A59/B84, p.197)

Again, we suggest to roughly interpret "the understanding thereby contradicts its general rules of thinking" as such a contradiction between different levels of reasoning.

Now, crucial for this kind of reasoning is the assumption that such reasoning chains can be recursively and coherently defined only for deductive inferences, but not for non-deductive ones. Otherwise one could question, e.g., modus ponens "without presupposing that it is true", or this or other inference rules "would be no a priori ground for the possibility of" reasoning, since there were other reasoning chains providing other grounds. So, the question is whether one can also justify non-deductive inferences which are not truth preserving along the lines of the schema in figure 8.1. And the answer is again quite destructive: *Yes, there are.* Haack (1976, p.115) provided an instance of such an inference rule, *modus morons*. The argument is as follows:

1. $q,\ p \to q \mid\!\vdash_1\ p$ (partial definition of $\mid\!\vdash_1$)
2. $T(q)$ (assumption)
3. If $T(p)$ and $T(p \to q)$, then $T(q)$. (from the definition of $\to$)
4. Hence: $T(p)$. (by modus morons of level 2, simplification, 1–3)

So, by applying modus morons on the meta level, one can prove that it is truth preserving (the assumptions in 2 allow for inferring 4), and hence justified. Although applying modus ponens on the meta level allows for proving that it is not truth preserving, this does not matter here, because modus ponens is not supposed to be part of this reasoning chain.

Note, that this time step 4 is in need of another inference rule which allows for inferring from $T(q)$ and $T(p \to q)$ that if $T(p \to q)$, then also $T(q)$ which is classically valid, but it can be questioned whether such an inference is valid in a *moronian logic* (see Hale 1978, p.112). However, we can circumvent this problem by framing the problem a little bit differently: Let us make the assumption of optimised justification explicit by help of the principle that truth preservation of an inference rule is sufficient for its justification:

$$TP(\vdash) \to J(\vdash)$$

Furthermore, let us consider TP to be fundamental in the sense that we know without any reasoning whether it applies to an inference rule or not—in fact we presuppose that it coincides with the valuation according to the ordinary truth tables, but we do not want and need to make this explicit. Modus ponens is truth preserving, i.e. $TP(\vdash_1)$, modus morons is not: $\neg TP(\mid\!\vdash_1)$. Note that TP serves the same role as the assumption about success served in infinitist inductive reasoning: induction was plainly supposed to be successful regarding the past, anti-induction was plainly supposed to be not successful regarding the past (we did not argue that by help of anti-inductive reasoning it turns out that anti-induction was successful in past). Now, the justification of level 1 modus ponens is simply:

1. $TP(\vdash_1)$

2. $TP(\vdash_1) \rightarrow J(\vdash_1)$
3. Hence: $J(\vdash_1)$ (by level 2 modus ponens and 1,2)

It is nice to note that anti-deduction ($|\vdash$) in the sense of an *anti-modus ponens* fails to provide a justification (in contrast to the coherentist and infinitist justification of anti-induction): Define: $p,\ p \rightarrow q\ |\vdash_1\ \neg q$; since $\neg TP(|\vdash_1)$ one cannot infer by level 2 anti-modus ponens a justification of it. However, anti-deduction in the sense of licensing the inference of the negation or "*de*-negation" of what deduction licenses (e.g.: if $\varphi \vdash_1 \psi$, then $\varphi |\vdash_1 \neg\psi$ and if $\varphi \vdash_1 \neg\psi$, then $\varphi |\vdash_1 \psi$) can be shown to be justified (as well as unjustified):

1. Define: $X\ |\vdash_1\ \neg p$ iff $X\ \vdash_1\ p$
2. $\neg TP(|\vdash_1)$
3. $TP(|\vdash_1) \rightarrow J(|\vdash_1)$
4. Hence: $TP(|\vdash_1)$ (by level 2 anti-deduction and 2)
5. Hence: $\neg J(|\vdash_1)$ (by level 2 anti-deduction and 3,4)
6. Hence: $J(|\vdash_1)$ (by level 2 anti-deduction and 5)

There is also a version of *modus morons* which allows for proving (of course in its own terms) its justification, without proving (in its own terms) that it is not justified:

1. Define: $\neg p,\ p \rightarrow q\ |\vdash_1\ q$
2. $\neg TP(|\vdash_1)$
3. $TP(|\vdash_1) \rightarrow J(|\vdash_1)$
4. Hence: $J(|\vdash_1)$ (by the so defined level 2 modus morons, 2,3)

Now, we have seen that some deductive rules can be justified this way, and some can not. An anti-modus ponens could not be justified this way, the version of modus morons in 1 could be justified. Note that according to their own terms, these inference rules are truth preserving (although they are not according to the terms of classical logic). So, according to their own terms, they are deductive inferences. That they all can be justified by such kind of reasoning might be also considered as a virtue than a vice, given logical pluralism (for a recent discussion discussion of the latter see Cohnitz, Pagin, and Rossberg 2014).

Even faced with this multitude of justified inferences, one can argue for the optimality of classical logic. So, e.g., Schurz (2018, sect.5.1) argues for this in the following way: Typically, if there is an alternative logic, then there is also a translation which allows to embed the alternative logic into classical logic:

> "I conjecture that a [...] translation strategy applies to all kinds of non-classical logics (even those not characterizable by finite matrices). My reason for this conjecture is that all non-classical logics known to me use classical logic in their meta-language in which they describe the semantics of their non-classical principles. Therefore there must exist ways to translate the principles of these logics into classical logic, by introducing additional operators into the language of classical logic corresponding to the semantical concepts of the non-classical logic (e.g., non-standard truth values in the case of many-valued logic)." (Schurz 2018, sect.5.1)

This is a clearly tenable strategy. However, it does not allow for arguing for a special status of classical logic (it ought to be noted that Schurz 2018, does not aim to argue for this). One might be also able to proof the optimality of non-classical logics: Once one starts to formulate, describe, and argue for a logic on the meta level in the same way as on the object level (some meta-mathematicians think, e.g., that mathematicians argue by help of intuitionistic logic, and hence suggest to also use intuitionistic logic on the meta level). Then also other—non-classical—logics can be shown to be optimal in this sense.

Another problem related to this is the following one: What, if a metalogic allows for a broader notion of *translatability* than classical (meta-)logic does? In such a case there is no longer a guarantee for embeddability of such a non-classical logic into classical logic, although embeddability might be possible the other way round. Hence, the optimality argument from above and its conjecture for classical logic needs to be considered as conditional on its current "past success". So, also in this sense a deductive justification of classical logic is in the same boat as a deductive (meta-inductive) justification of induction.

To sum up, there is indeed a parallel problem for an infinitist noncircular deductive justification of deduction as there is for such an inductive justification of induction: Both allow not only for the deductive/inductive justification of deduction/induction, but also for the deductive/inductive justification of anti-deduction/anti-induction. Haack (1976) provided an example which shows that such reasoning fails, if one allows for evaluating whether an inference rule preserves the truth, the very same inference rule is applied. Our examples show that even if we grant classical logic the evaluation of truth preservation (likewise as we granted classical logic the evaluation of success in the case of induction), such a justification still fails. The question is, can one do better by help of an inductive justification of deduction?

8.3 On an Inductive Justification of Deduction

Let us now come to the other horn of the dilemma, the inductive justification of deduction. Recall, "we cannot justify deduction inductively, because to do so would be, at best, to show that *usually*, when the premisses of a deductive argument are true, the conclusion is true too—which would be *too weak*" (Haack 1976, p.112). Now, if 'usually' means 'occasionally not', then induction is not too weak for justifying deduction. As Huber (2017, sect.7) argues, we can take as evidence all the known particular inferences that are structurally equivalent with classical deductive rules. From these inferences we know that they have at least one false premise or a true conclusion (so they do not occasionally lead from true premisses to wrong conclusions). Hence, by induction we are justified in assuming that all these inferences have at least one false premise or a true conclusion which means that classical deductive rules and all structurally equivalent rules are truth preserving, i.e. epistemically justified (see Huber 2017, p.528).

Now, at the beginning of the chapter we stated that the epistemic end of an inference is not only truth preservation, but guaranteed truth preservation, i.e. truth preservation *with certainty*. This addition is relevant, because otherwise we would fail in modal contexts and all inferences with contingently true conclusions would be licensed as deductive ones. The problem with an inductive justification of deduction is that we cannot account for truth preservation *with certainty*. Huber (2017) suggests to not "*blame*" induction for the impossibility of achieving truth preservation *with certainty*, but us: "it is not the principle of induction that is to be blamed. If anyone, it is *we* who are to be blamed, because *our* cognitive limitations prevent us from establishing the premise needed to inductively infer this stronger conclusion" (Huber 2017, p.528). The stronger premise concerns the known particular inferences that are structurally equivalent with classical deductive rules. However, we would need to know that *with logical necessity* they have at least one false premise or a true conclusion. Since we do not know this, we also cannot apply induction and infer that all these inferences have with *logical necessity* at least one false premise or a true conclusion which would mean that classical deductive rules are truth preserving *with certainty*.

We now want to outline an argument which allows also for an inductive justification of deduction via truth preservation *with certainty*. It is interesting to note that an infinitist deductive justification of deduction as discussed in the preceding section can be transformed into a finite inductive justification of deduction, simply by stopping the regress at some level and applying an inductive principle. Here is how this might be implemented: Assume we aim at justifying $\vdash_1$ (e.g. level 1 modus ponens). We do so by help of $TP(\vdash_1)$, $TP(\vdash_1) \rightarrow J(\vdash_1)$ and $\vdash_2$. So, the justification of $\vdash_1$ hinges on that of $\vdash_2$. Now, we go on with justifying $\vdash_2$, the meta level rule. We do so by help of $TP(\vdash_2)$, $TP(\vdash_2) \rightarrow J(\vdash_2)$ and $\vdash_3$, the meta meta level rule. Hence we receive $J(\vdash_2)$ on the basis of $J(\vdash_3)$. Analogously we base the justification of $\vdash_3$ on $\vdash_4$ ... that of $\vdash_{n-1}$ on $\vdash_n$. Hence, we get $J(\vdash_2), \ldots, J(\vdash_{n-1})$ based on $J(\vdash_n)$. Now, by induction we infer that all structurally equivalent inferences $\vdash_i$ are justified on the basis of $\vdash_n$, so also (not by help of modus ponens, but universal instantiation) that $\vdash_{n+1}$ is justified ($J(\vdash_{n+1})$) on the basis of $J(\vdash_n)$. Since $J(\vdash_{n+1})$ justifies $\vdash_n$ by the same scheme as above, we get $J(\vdash_n)$. And, hence, we have an unconditional inductive justification of $\vdash_1$. The schema of such a justification is presented in figure 8.2.

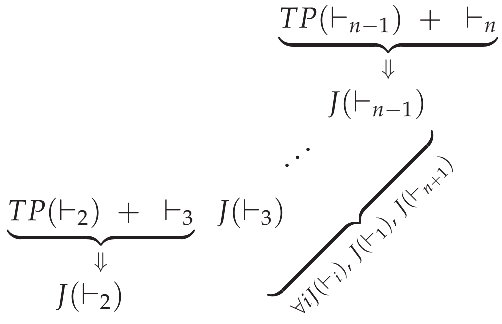
$\underbrace{TP(\vdash_{n-1}) \ + \ \vdash_n}$
$\Downarrow$
$J(\vdash_{n-1})$
$\cdot^{\cdot^{\cdot}}$
$\underbrace{TP(\vdash_2) \ + \ \vdash_3}$
$J(\vdash_3)$
$\Downarrow$
$J(\vdash_2)$
$\forall i J(\vdash_i), J(\vdash_1), J(\vdash_{n+1})$

Figure 8.2: Schema of transforming an infinitist deductive justification of deduction to a finite inductive justification of deduction.

Note that the deductive justification at each level is due to truth preservation *with certainty*, since for each level it follows by help of the deductive means of the higher level inference that the lower level inference is guaranteed to preserve the truth, and hence is justified. Now, by applying induction on this basis, we can infer that *with certainty* truth is preserved, and hence we gain an inductive justification of $\vdash_1$.

8.4 Summary of Main Results

In chapter 5 we provided a justification of induction. In section 8.3 we outlined a justification of deduction. Now, once we ask for a stronger notion of *justification* than coherence, we cannot achieve both justifications at one and the same time. From Hume's dilemma we went with the horn of providing a deductive justification of induction. From Haack's dilemma we went, roughly speaking, with the horn of providing an inductive justification of deduction. Now, by combining deduction and induction, a "single horn" in form of an impossibility remains: We cannot justify deduction *and* induction by help of deduction *and* induction, because such a justification would be *circular*. However, this should not bother us further: Granting any form of inference is an *a priori ground for the possibility of reasoning at all* might Kant have said and we could not agree more (although one might express this a little bit less prosaic: *Nope, sorry!*).

We now want to outline how one could characterise the different inference methods by help of their different ways of optimising success (in the long run): We know that the main epistemic ends for which the different inference methods are assumed to be adequate means are as follows: Deduction is a sufficient means for the epistemic end of truth preservation (also in the short run). (Meta-)Induction is a sufficient means for the epistemic end of *truth*-based relative success preservation in the long run. Abduction is a sufficient means for the epistemic end of *data*-based relative success preservation in the long run.

Now, let us spell out the deductive feature of truth preservation in terms of success: Let $Y_1, \ldots, Y_{t-1}$ be the premisses of a deductive inference, and let Y_t be the conclusion of such an inference. Let $y_1, \ldots, y_{t-1}$ and y_t be the respective truth values: They could be binary ($\{0,1\}$), they could be k-ary ($\{k_1 \in [0,1], \ldots, k_m \in [0,1]\}$ for some $k_1, \ldots, k_m$), they could be degrees of belief ($[0,1]$). In the simple case of $n = 2$, i.e. with Y_1 as premise and Y_2 as conclusion, "truth" preservation amounts to the constraint:

$$y_2 \geq y_1$$

Note that this is clearly satisfied for the case of classical logic, but also for the case of probabilism due to the consequence theorem. It is not necessarily satisfied for the case of many valued logic, since many rules of these systems are intended for the preservation of so-called *designated values*, i.e. values that are considered to be relevant for validity of inferences in such systems. In principle this allows for inferences where $y_2 < y_1$ as long as y_2 is a designated value, if y_1 is such a value. However, for simplicity reasons we restrict our considerations only to rules satisfying $y_2 \geq y_1$. In the case of $n > 2$, "truth" preservation amounts to the constraint:

$$y_t \geq \prod_{u=1}^{t-1} y_u$$

Again, this clearly holds for classical logic: Whenever all premisses have value 1, then also the conclusion has value 1. And by assuming probabilistic independency of the premisses $Y_1, \ldots, Y_{t-1}$, it holds also for probabilism: If $Y_1, \ldots, Y_{t-1} \vdash Y_t$, then $Pr(Y_t) \geq Pr(Y_1 \& \cdots \& Y_{t-1}) = \prod_{u=1}^{t-1} Pr(Y_u)$. Again, also for many valued logic this does not hold generally (but, e.g., for the typical rules for & and $\vee$ taking the minimal or maximal value). For simplicity reasons we restrict our considerations only to cases with probabilistically independent premisses etc.

We can put forward reaching the truth or certainty 1 as the absolute epistemic end. Analogous to before, we claim that it is the task of the deductive method f_d to infer Y_t, and $\ell_{d,t}$ measures somehow the distance between the inferred conclusion and the truth. We define $s_{d,t} = 1 - \ell_{d,t}$ as a measure for the score of the inference of f_d. If we think of f_d having inferred $Y_1, \ldots, Y_{t-1}$ before, then we might interpret $\ell_{d,u} = 1 - y_u$ $(1 \leq u \leq t)$ as the losses of f_d in deducing Y_u, and $s_{d,u} = y_u$ as the respective scores. Now, by geometrically averaging the scores of the premisses, we define a geometrical success measure for f_d, the deductive method's success in deducing the premisses for Y_t:

$$succ_{d,t-1} = \left(\prod_{u=1}^{t-1} s_{d,u} \right)^{1/(t-1)}$$

Keep in mind that this holds for a more general than only a binary setting (not necessarily $succ_{d,t-1} \in \{0,1\}$). Given the restricted notion of truth preservation we can see that deductive methods aim at and allow for the *preservation of success* in the sense of scoring:

$$s_{d,t} \;\geq\; succ_{d,t-1}$$

We can compare this with the optimality result of meta-induction, stating that in the long run an inductive method f_i will be at least as successful, as any other method f_b is: $\lim_{t\to\infty} succ_{i,t} \geq succ_{b,t}$. This means that in the long run, (meta-)inductive methods aim at and allow for the *preservation of relative success* in the sense of *truth based scoring* (where f_b is any accessible alternative method):

$$s_{i,t} - s_{b,t} \;\geq\; succ_{i,t-1} - succ_{b,t-1}$$

Finally, in chapter 7 we have transferred this optimality result to abduction, where success is no longer truth, but data based, and data might contain noise and error. So, also abductive methods f_a aim at and allow for the *preservation of relative success*, but this time not in the sense of truth based, but in the sense of *data based scoring*:

$$s_{a,t} - s_{b,t} \;\geq\; succ_{a,t-1} - succ_{b,t-1}$$

In terms of success, we can express the main difference between deductive and inductive (in a wide sense) inferences as follows: Deductive inferences aim at absolute success, whereas inductive inferences aim at relative success. This means that the conclusion of a deductive inference can never be farther from the truth than the premisses are. Figure 8.3 illustrates this fact. In contrast, the conclusion of an inductive inference might be farther away from the truth than the premisses are. However, *in comparison* with accessible alternative inference methods, i.e. relative to such inference methods, the conclusion will not be farther from the truth than the premisses are (in the long run). This fact is illustrated in figure 8.4.

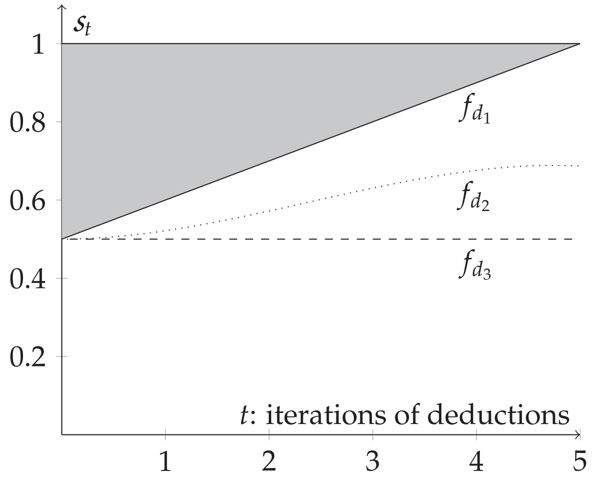

Figure 8.3: Example of different deductive methods and their iterated application to different sets of premisses: f_{d_1} will reach the truth. The application of deductive method f_{d_3} will not lead to the truth, but wont deviate from it either. Similarly for the application of f_{d_2}. In deductive inferences the distance from the truth will never grow (e.g. the shaded area between the truth and f_{d_1} remains constant or shrinks, but never grows). Note that 'deduction' is understood here not only in terms of classical (two-valued logic), but also in terms of relations of probability conservation (any inference between logical consequences which preserves the probabilistic value is in this sense similar to a classical deductive inference).

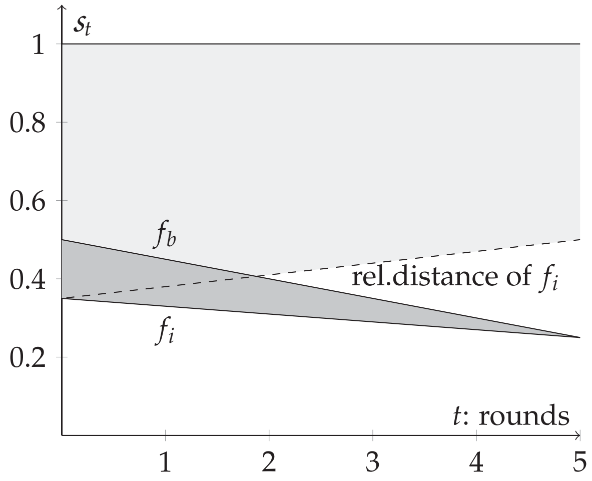

Figure 8.4: Example of the development of scores of a meta-inductive method f_i: As one can see, in this example with increasing number of rounds the method's distance from the truth grows. However, compared to the accessible object method, f_b, the relative success grows: The relative difference of the scores (which is ascended) is marked by the dashed line. Note that this growth holds only while approaching the limiting case.

Part III

Optimisation in the Social Epistemic Realm

Chapter 9

Testimony

In this chapter the problem of testimony is formulated and traditional accounts are briefly discussed. Afterwards, it is outlined how meta-induction can be employed for a reliabilistic testimony acceptance practice. Finally, the general case of expert testimony is investigated.

Classical epistemology focused mainly on the *epistemic individual* with its individual notion of *belief* and *knowledge*, as well as the *justification* thereof. Although epistemologists and philosophers of science discussed social phenomena from time to time and in scattered contexts, it took until the mid 1980s when a social turn in epistemology bloomed, grew fast, and started to set forth new topics on the agenda of epistemology. Considering not only an individual epistemic agent, but a set of such agents, puts forward new problems which are not covered by solutions of the classical realm. Classically, the focus was on individual sources of belief and knowledge, i.e. *perception*, *introspection*, *memory*, *reasoning*. In part II we focused mainly on justifying our reasoning practices. In the social domain another source of belief and knowledge is predominant, namely *testimony*. One of social epistemology's first and perhaps most influential proponents, Goldman, puts it this way: "What others think is part of one's total evidence, a social part of that evidence" (Goldman 2011a, p.16).

Looking back at about forty years intensive research since the social turn, Goldman (2011a) suggests a tripartite division of the domain of social epistemology: (i) there are individual epistemic agents with social evidence, (ii) there are collective epistemic agents, and (iii) there are epistemic systems where individuals and collectives are embedded in. (iii) concerns system design and is perhaps the most holistic social epistemological approach, ranging from science itself, over law, democracy, to education as the main modules of our societies (see Goldman 1999, prt.3). Here we concentrate on the more encapsulated topics (i) and (ii). The main topics of (i) concern the problem of testimony and peer disagreement. The problem of testimony consists in the task to identify and justify testimony acceptance practices. There, e.g., in particular the problem of the case where an expert testifies to a novice shows up. But also the problem of how to deal with disagreeing testimonies of peers resulted in a comprehensive debate. The main topic of (ii) is the problem of judgement aggregation and, to our understanding, also the problem of the wisdom of the crowds. Judgement aggregation becomes especially relevant, when a collective of epistemic individuals has to act as one single unit, e.g., when a jury has to render *a* judgement or when a panel has to provide *a* recommendation (a prominent alternative to collective group decision by aggregation is that of finding a consensus via deliberation—for an investigation of truth-conduciveness of deliberation see Hartmann and Rafiee Rad 2018). Early impossibility and characterisation results of the 1960s attracted lots of research in social choice theory and neighbouring disciplines and triggered a lively debate which also entered the domain of social epistemology and is today one of the most prominent areas of research in this field. In the course of judgement aggregation, problems of the wisdom of the crowds become also urgent: Whenever a collective acting as a single unit performs better than an average individual, one calls such an act a *wise crowd effect*. The problem of the wisdom of the crowds concerns the question under which conditions which forms of (aggregated) collective action produce such a wise crowd effect. Perhaps most famous is the so-called *Condorcet jury theorem*.

Now, it is interesting to note that already very early on research on judgement aggregation was linked to the investigation of wise crowd effects: In the same work where the famous *jury theorem* was proven first, also one of the first impossibility results of judgement aggregation, known today as *Condorcet paradox*, was discussed, namely the problem that majority aggregation of preferences (orderings) might be incoherent in the sense that the majority preference (ordering) might produce a circle and hence becomes intransitive (due to $a > b, b > c, c > a$ one has to exclude $a > b, b > c \rightarrow a > c$).

In this part of the book we are going to apply the theory of meta-induction to the domain of social epistemology ((i) and (ii)). Our investigation is mainly conceptual. For simulations of meta-inductive performance in the social realm we refer the reader particularly to (Schurz 2009, 2012b). For a comparison of meta-induction with other social strategies see (Schurz and Thorn 2016). We discuss the problem of optimisation with respect to these problems in the social epistemological realm: testimony (this chapter), peer disagreement (chapter 10), and judgement aggregation (chapter 11). Finally, we link the main theoretical property of the framework introduced in part I (convexity of the loss function ℓ) to assumptions made in the debate of wise crowds (chapter 12).

We begin our investigation of meta-induction in the social realm by studying its application to the problem of testimony: In section 9.1 we formulate the problem of testimony. In section 9.2 we provide an overview of traditional approaches, and in section 9.3 we outline our meta-inductive approach. Finally, in section 9.4 we discuss problems related to the case of experts testifying to novices.

9.1 The Problem of Testimony

A great bulk of what we know depends in some way or other on the testimony of others. Whatever proposition p you think of, whenever you belief or disbelief p, know that p is the case or know that p is not the case, and sometimes even when you neither belief nor disbelief p, almost always testimony will be involved as a source of your belief, disbelief or suspension of judgement. You might have read a report, article, textbook, listened to your teacher, supervisor, a friend, family, news etc.: Most probably your epistemic attitude towards p is influenced by someone testifying p or $\neg p$.

One of the main features of testimony seems to be that "communication is an efficient mode of increasing knowledge because information transmission is typically *easier, quicker, and less costly than fresh discovery*. [...] Since not every member of a community *observes each fact other members observe*, there is room for veritistic improvement through communication" (see Goldman 1999, p.103). Clearly, since testimony has *de facto* such a central role in our forming of epistemic attitudes and producing knowledge, the question about its rationale arises. Is testimony an optimal means towards some epistemic end as, e.g., truth? Given an informal characterisation, this seems to be not the case: Informally, testimony is often characterised as an utterance, a speech act, where the testifier (intentionally) aims at conveying information:

> "Testimony of the informal kind—roughly, *saying something in an apparent attempt to convey (correct) information to someone else*—plays a very large role in our lives and raises the question of the importance of testimony for knowledge and justification." (Audi 2011, p.150)

Now, even if one puts forward as a necessary condition that the testifier aims at conveying correct information, this neither implies that in fact correct information is *conveyed correctly* (for crazy communication chains think, e.g., of the children's game *telephone*), nor that the correctly conveyed information is in fact *correct information*. So, clearly, testimony is not truth conducive in general. How then can it be an adequate source of knowledge? This is the problem of testimony we are going to address in the following sections.

In our formal analysis we will not make any assumption about the intentions of the testifier. So, she might be aiming at conveying correct information, but she might be also aiming at deceiving. Furthermore, we also make no assumptions about the circumstances of the speech act and what counts as part of such a speech act. One might interpret Kant's famous example of packing your bags in front of someone else in order to deceive her into thinking you are going on a trip as a case of testimony or not (see Kant 1762/1997, 27:447, p.202). Rather, we only assume that whoever counts as testifier and whatever counts as testimony, we can attach to the testifier somehow a measure of how successful she was in testifying in past.

In the next section we outline traditional and modern approaches to the problem of testimony. We characterise some general features of testimonial practice and argue that one approach, namely reliabilism, seems to allow for the best explanation of the epistemic rationale of testimony. However, the reliabilist solution to the problem of testimony is externalistic, and hence incorporates also the problems of externalism. For this reason we will afterwards outline an internalistic reliabilist alternative in section 9.3 and indicate how this alternative allows for an optimality justification of testimony via meta-induction.

9.2 Traditional Approaches

We have seen that the main problem of testimony consists in answering the question if, and if so, how testimony can be considered as an *adequate* epistemic source of knowledge. In the epistemic tradition three major approaches can be distinguished:

- Testimony is no adequate source of knowledge (Descartes; consideration of Descartes in the context of testimony due to (Zollman 2014)).
- Testimony is *a priori* an adequate source of knowledge (Thomas Reid).
- Testimony is *a posteriori* an adequate source of knowledge (Hume).

As we will see soon, a posteriori justifications of testimony are in general reductionistic:

> "Besides the word of the speaker, hearers also causally depend in believing testimony on other fundamental sources of knowledge like perception, memory, learning, and inference. *Can the reliability of testimony be justified by appeal to these sources?* This question represents the dominant epistemological problem of testimony—is testimony an autonomous source of epistemic authority?" (par.1 Adler 2012)

But let us first provide some evidence that these authors can be also attached to the mentioned positions.

We begin with Descartes who claims in his *Discourse on Method*, *Rule III*:

> "In the subjects we propose to investigate, our inquiries should be directed, *not to what others have thought*, nor to what we ourselves conjecture, but to what we can clearly and perspicuously behold and with *certainty deduce; for knowledge is not won in any other way*." (Descartes 1975, Rule III, p.5)
>
> and:
>
> "And thus I thought that *book learning*, at least the kind whose reasonings are merely probable and that do not have demonstrations, having been composed and enlarged little by the opinions of many different persons, *does not draw nearly so close to the truth as the simple reasonings that a man of good sense can naturally make about the things he encounters*." (Descartes 1637/1998, part one, p.7)

Descartes' position on testimony fits perfectly well with his high standards for epistemic justification in general. For practical reasons we might reasonably incorporate beliefs via testimony, but in theory one's beliefs have to be grounded on one's own fundamental basis (as, e.g., the *cogito*, and not a *cogitas*).

Reid states in his *Inquiry*, section XXIV: *Of the Analogy Between Perception and The Credit We Give to Human Testimony*:

> "In the testimony of Nature given by the senses [i.e.: perception], as well as in human testimony given by language, things are signified to us by signs." (see Reid 1764/1785/1788/1983, p.90)
>
> and:
>
> "The wise and beneficent Author of Nature, who intended that we should be social creatures, and that we should receive the greatest and most important part of our knowledge by the information of others, hath, for these purposes, implanted in our natures *two principles that tally with each other*. [. . .] The first of these *principles is a propensity to speak truth* [. . .] Another original principle implanted in us by the Supreme Being is, *a disposition to confide in the veracity of others*, and to believe what they tell us. This is the counterpart to the former; and, as that may be called *the principle of veracity*, we shall, for want of a

> more proper name, call this *the principle of credulity*." (see Reid 1764/1785/1788/1983, pp.94f)

Now, whereas Descartes was quite pessimistic or sceptical regarding the adequacy of testimony, Reid was much more optimistic and non-sceptical—one might also consider it a bit epistemically *naïve*. However, we think this would fall short of a benevolent interpretation. Rather, one can put it differently and draw a parallel to approaches to the epistemic problem of justification we have discussed at length in part II in this book: Whereas Descartes grants justification only to beliefs and sources of beliefs which accommodate high epistemic standards, Reid seems to grant justification of beliefs and sources of belief—in particular the source of testimony—*per default* and just withdraws it in case there is a defeater. Putting it this way brings it in parallel to the discussion of the justification of induction and the falsificationist response of failing to provide reasons for an epistemic stance. Furthermore, as we will see below, seeing it this way also links Reid's position closer to modern approaches of the so-called *interpersonalist camp*.

Finally, let us provide some evidence for our short characterisation of Hume's position: In his *Enquiry*, *Of Miracles* he writes:

> "We may observe that there is no species of reasoning more common, more useful, and even necessary to human life, than that which is derived from the *testimony* of men, and from the reports of eye-witnesses and spectators. [...] *Our assurance* in any argument of this kind *is derived from no other principle than our observation of the veracity of human testimony*, and of the usual conformity of facts to the reports of witnesses." (see Hume 1772, p.127)
>
> and:
>
> "The reason why we place any credit in witnesses and historians, is *not derived from any* connexion*, which we perceive* a priori*, between testimony and reality*, but because we are accustomed to find a conformity between them." (Hume 1772, p.129)

In contrast to Descartes, Hume considers testimony clearly as an adequate (even necessary) source of belief and knowledge. However, contrasting Reid, he explicitly declines to argue for the adequacy of this source by help of an a priori principle (implanted, put into effect by some supreme being). Rather, he is after an a posteriori justification, an inductive one. So, given the pairs of notions *a priori/a posteriori* and *acceptance/denial*, we can distinguish four positions, as depicted in table 9.1.

	denial	*acceptance*
a priori	Descartes: Testimony does not suffice high epistemic standards of justification.	Reid: Testimony is justified due to *veracity* and *credulity* (implanted by God).
a posteriori	Hume: Testimony is unjustified in case of unreliable agents.	Hume: Testimony is justified in case of reliable agents.

Table 9.1: Traditional positions in the debate on the epistemic justification of testimony

Note that there is a connection between these positions: Hume's position and that of Descartes coincide in case of unreliable testifiers. Descartes' approach can be also considered as a special case of Hume's position where all testifiers are always unreliably (e.g. according to high epistemic standards). And Reid's position can be considered as a special case of Hume's approach where all or most of the testifiers are in general reliable. It is interesting to note that empirical studies suggest some rudimentary implementation of the principles mentioned by Reid: Studies of signalling (birds, apes, etc.) suggest some rough "innate propensities of" veracity and credulity in the following sense (see Goldman 1999, p.106):

- More alarm calls are produced to conspecific audiences than to audiences of another species.
- Fewer alarm calls are produced when there is no audience at all.

So, again, Reid's position is less naïve as one might have thought and a weaker and naturalised version of justifying testimony per default rules might be perfectly reasonable.

We now want to provide a simple model for these positions. We can do so by applying probabilism and the Bayesian framework. Within this framework the explication of the traditional positions is quite easy: We can describe them simply as different ways of updating in the case where testimonial evidence enters the scenery. Let us label cases where someone testifies p with '$Test_i(p)$'. In the Bayesian setting, the problem of testimony amounts to the question of how to incorporate testimonial evidence. So, assume Pr_i are one's degrees of belief *prior* acquiring such evidence, and Pr_u are one's degrees of belief *posterior* acquiring testimonial evidence (after *u*pdating). According to Bayesian orthodoxy, once one receives new evidence $Test_i(p)$, one needs to updated by the Bayesian rule:

$$Pr_u(p) = Pr_i(p|Test_i(p))$$

Now, the three positions from above amount to the following updates in the Bayesian setting: Descartes' testimonial scepticism states that testimony should not have any influence at all, hence:

$$Pr_u(p) = Pr_i(p) \tag{TDescartes}$$

Note that this means that an independency between facts and testimony is already hard coded in the priors (i.e. a priori): $Pr_i(p|Test_i(p)) = Pr_i(p)$. Reid's testimonial acceptance rule per default states that p should be accepted, once someone testifies p:

$$Pr_u(p) \approx 1 \tag{TReid}$$

Also this means that the impact of testimony is hard coded in the priors, that a strong correlation is already fixed in the priors (i.e. a priori): $Pr_i(p|Test_i(p)) \approx 1$. Finally, Hume's account amounts to incorporating the testimony in accordance with the testifier's reliability rel_i (rel_i is intended to measure how good a testifier agent i is, i.e., e.g. in a binary testifying task, how often her testimony that some proposition p holds was correct; we will provide more details on and an exact characterisation of rel in a minute):

$$Pr_u(p) \propto rel_i \tag{THume}$$

Clearly, the Bayesian framework has it that also the Humean update is already somehow wired in the priors. However, whereas in (TDescartes) and (TReid) p is relevantly conditionalised on $Test_i(p)$ directly, the idea of the Humean account is to conditionalise p relevantly on $Test_i(p)$ *and* other evidence about i testifying some propositions in other contexts and matches or mismatches of these testimonies with the facts in the other contexts. In this sense also in the Bayesian setting (THume) can be considered as an a posteriori approach to testimony (namely incorporation of testimony posterior grasping reliability information).

Now, let us come to an evaluation of these testimony acceptance practices. Clearly:

> "The veritistic merits of a hearer acceptance practice cannot be assessed in isolation from the reporting practices that it complements. This point can be appreciated by reflecting on results from game theory. A particular strategy for playing a certain game can be very successful when pitted against a second strategy but much less successful when used against others." (Goldman 1999, p.109)

So, if we consider, e.g., a testifying practice that generates only truths, then blind trust in the sense (TReid) does the job of achieving truths best, whereas (TDescartes) cannot employ testimony as epistemic source at all and (THume) will at some point in time employ the testimony, but has some initial costs of evaluating the reliabilities. On the other hand, if we consider a testifying practice that generates only falsities, then (TReid) performs worst (however, at some point in time it will withdraw its default acceptance of the testimony), whereas (TDescartes) is not harmed at all, and (THume) will benefit perfectly since it proportions its degrees of belief according to the reliability of the testifier and once it is low, $Pr(\neg p)$, which is indirectly proportional to the reliability, will get high.

Prima facie, it seems that in comparison with the other approaches to testimony, (THume) is better off: (TDescartes) cannot account for the *de facto* success of testimony and fails in testimony-affine environments. (TReid) can account for the *de facto* success of testimony by reference to a priori principles, perhaps also naturalised variants thereof, however, it fails in testimony-averse environments. Finally, (THume) can also account for the *de facto* success of testimony by reference to a posteriori principles and neither fails in testimony-affine nor testimony-averse environments. What is more, it can be justified by applying induction. The argument in its qualitative form for reliable agents is as follows: If testifier i was correct in her past testimonies, then she will be correct in her present testimony (by induction). Testifier i was correct in her past testimonies (observation). Hence, i will be correct in her present testimony—for which reason one should accept it in accordance with (THume). The argument for unreliable agents is based on the same principles and goes analogously.

Furthermore, note that (THume) allows also for accounting for a feature of testimony which has been employed famously for a very long time now. The feature we are speaking about is the veritistic value of *independent, but coinciding testimonies*. It is not by accident, e.g., that Catholicism stresses independency of the claims in the four Gospels, or that historians stress the independency of historical reports, or that judges, prosecutors, and lawyers are after independent eye witness reports in order to argue for their main theses. Rather, this is standard methodology. The rough idea is that some proposition p can be considered as true, if it is testified independently of each other by several testifiers. The more improbable it is that two testifiers came up with a testimony on p, given p is true, and the more such testifiers there are, the more plausible we consider p. If, e.g., we think of a huge set of possible conceivable hypotheses $h_1, \ldots, h_{1.000.000}$ to explain some fact, and we have a high number n of testifiers claiming that h_{22516} is the case: $Test_1(h_{22516}), \ldots, Test_n(h_{22516})$, then it seems quite improbable that they all came up with the same hypothesis, once we assume that they are at least partly independent of each other. A simple Bayesian analysis reveals why this is so. According to the odds version of Bayes' theorem, the posterior odds equal the likelihood ratio times the prior odds. If we compare the case with one testimony $Test_1(p)$ with the case of two (or more) testimonies $Test_1(p), Test_2(p)$, a boost by more testimonies would amount to an increase of the posterior probability:

$$\begin{array}{ccccc}
\overbrace{\frac{Pr(p|Test_1(p))}{Pr(\neg p|Test_1(p))}}^{\text{posterior odds}} & = & \underbrace{\overbrace{\frac{Pr(Test_1(p)|p)}{Pr(Test_1(p)|\neg p)}}^{\text{likelihood ratio}}}_{①} & \cdot & \overbrace{\frac{Pr(p)}{Pr(\neg p)}}^{\text{prior odds}} \\
\underset{?}{\wedge} & & & & \\
\frac{Pr(p|Test_1(p)\&Test_2(p))}{Pr(\neg p|Test_1(p)\&Test_2(p))} & = & \underbrace{\frac{Pr(Test_1(p)\&Test_2(p)|p)}{Pr(Test_1(p)\&Test_2(p)|\neg p)}}_{②} & \cdot & \frac{Pr(p)}{Pr(\neg p)}
\end{array}$$

Under what conditions does the inequality hold, does further testimony provide a boost? As one can see, the prior odds cancel out. So, what is relevant for the inequality are the likelihood ratios. It turns out that the ratios ① and ② are equal, if testifier 2 is blindly following testifier 1 in the sense that (see Goldman 2011b, p.122):

$$Pr(Test_2(p)|Test_1(p)\&p) = Pr(Test_2(p)|Test_1(p)\&\neg p) = 1$$

And ① $<$ ②, if the testimonies on a proposition are positively related to the truth of the proposition, and this positive relevance is invariant under conditionalisation on the other testimonies, i.e.:

$$Pr(Test_i(p)|p) > Pr(Test_i(p)|\neg p)$$

and this holds independently of conditionalisation on other testimonies:

$$Pr(Test_i(p)|p, Test_{j\neq i}(p)) > Pr(Test_i(p)|\neg p, Test_{j\neq i}(p))$$

So, given independent testimonies 1 and 2, Bayesian update allows for probabilistic boost (for a more detailed analysis of the Bayesian approach to multiple testimonies see Bovens and Hartmann 2003, sect.3.3).

Two provisos of the advantages of (THume) are in place. First, regarding the explanation of the *de facto* success of testimony as a social source of belief and knowledge: (THume) can explain this by reference to the *de facto* reliability of testifiers. However, very often we lack such information about the reliability of a testifier, for which reason the success of the individual acceptance practice remains unexplained. So, e.g., children learn a great bulk via testimony, although at an early stage they seem to have not really a system of reliabilities established. Also in the case of experts and novices, it is hard to see how novices could estimate the reliability of experts without being an expert—we will discuss problems related to this in section 9.4. In general, any reductionistic account to testimony has it that the reduction basis might be too weak in order to be able to explain the success of testimony. For this and other reasons the so-called *interpersonal view* emerged, according to which testimony is granted via an interpersonal relationship between the testifier and the recipient of the testimony. However, such alternative accounts are also prone to problems. Prominent is, e.g., the following dilemma (see Lackey 2008, 2011): Any such view needs to be *genuinely interpersonal* (to account for success) and "*epistemologically potent*" (to account for justification). However, if such an account is *genuinely interpersonal*, then it is "epistemologically impotent", since a *genuinely* interpersonal relation like trust allows not for truth veracity. In section 9.4 we will consider problems related to trust in more detail.

Second, (THume) is successful in testimony-affine as well as testimony-averse environments, however, there are also testimony-adversarial environments where testifiers fail, once they have high reliability, and testifiers succeed, once they have low reliability. In such environments (TDescartes) as well as (TReid) are better off. Since the performance of a testimony acceptance practice varies with the environment, one might wonder whether another epistemic end can be satisfied by an acceptance practice. For this purpose, Goldman (1999, p.110) suggests to "seek a veritistically *good* practice, even if it is not the *best* practice relative to this or that reporting environment, [where . . .] a good practice is one that produces veritistic improvements on average, over a range of actual and possible applications."

Now, as we have indicated in section 5.3 with a binary version of the so-called *no free lunch theorem* of online learning, averaging over all possible cases of event series (testimonies on p and outcomes of p) allows for no discrimination of such an acceptance practice, if no restriction to the series is put forward: Considering all possible series, there are equally many in which, e.g. (TDescartes) succeeds/fails as there are in which, e.g., (TReid) succeeds/fails. Similarly for (THume). However, if one puts forward a restriction, then one can discriminate among them.

Goldman (1999) suggests the following restriction: The reliabilities of the testifiers are not only subjective ones, but objective ones. I.e., rel_i captures to objective reliability of the testifier to testify correctly. So, only series are to be considered which are in accordance with rel_i. Now, in the simplest case we could take as a measure of reliability $rel_i = Pr(p|Test_i(p))$ which is sometimes also called a *measure for truth indication*: How good is i's testimony that p an indicator for the truth of p (see List and Pettit 2011, sect.4.1). Then, by the assumption that rel_i is an objective measure, this means that rel_i is not only about the recipient of the testimony's *estimation* of how good i testifies, but it is about how good i testifies in fact (in all relevant series). But then, clearly accepting $Test_i(p)$ in accordance with rel_i grants (THume) veritistic improvement on average, since $Pr_u(p) = Pr_i(p|Test_i(p)) = rel_i$, and hence, whenever $Pr_u(p) > Pr_i(p)$, then also $Pr_i(p|Test_i(p)) > Pr_i(p)$, and $Pr_i(p|Test_i(p)) > Pr_i(p)$ expresses a fact of the (average) event series under consideration. So, the "veracity of the testimony" is transmitted to accepting the testimony.

Now, to assume that rel_i is objective and expresses $Pr(p|Test_i(p))$ is not common practice. Rather, quite often, e.g., at court, it is practice to assume that rel_i is somehow objective, but expresses $Pr(Test_i(p)|p)$, the so-called *truth tracking reliability* of testifier i (see List and Pettit 2011, sect.4.1): This measure is intended for answering the question of how good a truth tracker i is. So, e.g., if a prosecutor tries to compromise an antagonist testifier by bringing forth some conflicts of interests, etc., it seems that the lawyer intends to show that although if p is or were the case, i does or would not testify p, because of a conflict of interest—so i is (objectively) unreliable. However, also such a measure of reliability (THume) is proven to increase veracity on average, since this property is preserved in applying the Bayesian formula (see thrm.4.1 Goldman 1999, p.121).

Now, what the objective reliability rel_i of testifier i is, is not accessible to us. In this sense the solution proposed by Goldman (1999) is an externalist one: If, from *God's eye view*, i is reliable, then we are justified in accepting her testimony, if not, then we are not. However, also, if from *God's eye view* $rel_i = 0$, then (TDescartes) is justified, and if from such a view point $rel_i = 1$, then (TReid) is justified. So, as long as we have no grip on rel_i, also these alternatives might be considered as justified. For this reason we think that one should also aim at an internalist approach to the problem of testimony which gets some grip on rel_i. And we think that the theory of meta-induction can serve this task.

In stating the epistemic end of veracity increase on average, Goldman (1999, p.110) asked also whether:

> "there [is] any acceptance practice that is optimal in all reporting environments, in other words, better in each reporting environment than every other acceptance practice would be?"

His answer was:"As in game theory, the answer appears to be "no."" And true, this answer is correct. However, with a slight modification of the epistemic end mentioned in the quote, one can provide a positive answer: If one is after an acceptance practice that is optimal in all reporting environments *in the long run* compared to every other *accessible* practice, then yes, there is an epistemic means to achieve this end. In the next section we outline the meta-inductive generalisation of (THume) in order to account for the problem of testimony.

9.3 The Meta-Inductive Approach

Now, what is a testimony acceptance practice that is optimal in all environments *in the long run* compared to every other *accessible* practice? Stating the epistemic end this way almost automatically refers to meta-induction.

In this section we outline a meta-inductive approach to the problem of testimony. Since we consider the problem of epistemic peer disagreement as a special case of the problem of testimony (both peers testify to each other), we sketch here only the meta-inductive solution and spell out the details of this approach in section 10.3 where we investigate the case of epistemic peer disagreement.

Testimony is a social source of knowledge. For our solution we need to shift the problem of justifying testimony as a practice for belief or knowledge acquisition to a fully social level, meaning that we assume there are testimonies from several individuals which might or might not conflict with each other. In our setting the task of the recipient of the testimonies, call her o, consists in incorporating these testimonies (accessible practices) in such a way that in the long run her reliability is optimal in the sense that her acceptance practice is not outperformed by any other practice. This means that in the long run either rel_o converges with rel_i of the best practice i or even outperforms rel_i. Now, once we understand reliability rel in terms of predictive success (see $\mathcal{succ}$ of section 2.3), then meta-induction is exactly such a testimony acceptance practice we are looking for, since it allows for relative learnability, i.e. optimality. The argument is as follows:

1. Frame testimonies in a prediction game G with truth $\mathcal{Y}$ and prediction practices $\mathcal{F}$.

2. Then we can define a measure for success $\mathcal{succ}$ (for details see definition 2.11).

3. Now, we consider the individual predictions $f_{i,t}$ also as testimonies (quantified versions of $Test_i(p)$).

4. We use $\mathcal{succ}$ to define rel as: $rel_{i,t} = \mathcal{succ}_{i,t}$.

5. We define a meta-inductive learner f_o based on $\mathcal{succ}$ (for details see definition 3.39 on f_{ami}).

6. This testimony acceptance practice f_o is long run access optimal (this follows from theorem 3.40) in the sense that

$$\lim_{t \to \infty} \mathcal{succ}_{o,t} - \mathcal{succ}_{i,t} \geq 0$$

   if we assume that ℓ which underlies $\mathcal{succ}$ is convex.

7. Hence:

$$\lim_{t \to \infty} rel_{o,t} - rel_{i,t} \geq 0$$

This is only a sketch of the meta-inductive approach which is structurally equivalent to our solution of the problem of epistemic peer disagreement. For how to exactly flesh out the definitions see section 10.3.

Let us first come to the advantages of this approach: *rel* is purely success based and in this we need not assume that it represents a general objective feature of the prediction series under investigation. Rather, *rel* is only about the past performance of the testifiers, hence, data which is in principle accessible to us. In this sense we consider the notion of reliability *rel* as internal. Furthermore, due to the long run optimality of the testimony acceptance practice f_o, we gain epistemic justification for reliability based testimony acceptance.

Now, let us also mention some provisos of this approach: First of all, f_o is guaranteed to be optimal not for any specific round t, but only in the long run. Second, f_o fully depends on the testimonies of others. So, what we have shown is that some reliabilist testimony acceptance practice is justified, but not all such practices. In particular, our approach disregards such acceptance practices which relevantly use own (prior) estimations in incorporating testimony. Furthermore, also the internalist version of *rel* seems to be a too weak basis for explaining the success of reliability based testimony acceptance: Children and novices seem to have only very little information about past success available. In general, also experts seem to lack such information: Think, e.g., on the case of a judge or prosecutor who needs to estimate the reliability of a testifier. In such a situation it is very often hard to figure out relevant information or come up with a testimonial track record. Clearly, in all these cases the approach outlined above cannot be employed directly for justifying reliability based testimony acceptance as an optimal means to achieve the stated epistemic end. However, the approach from above allows for justifying at least some such acceptance practices, and it might be considered as an idealised model which we apply to less ideal situations for the purpose of approximation.

Finally, one note about *accessibility* is in place. In framing the problem of accepting testimony within the setting of prediction games, we switched from comparing testimonial acceptance practices to comparing prediction practices with such a testimonial acceptance practice. Recall, the initial question of Goldman (1999, p.110) was about showing that a testimony acceptance practice is "better in each reporting environment than every other *acceptance practice*". Our framing of the problem compares such a testimony acceptance practice with any *accessible prediction practice*. This is quite a modification. However, the meta-inductive framework allows also for comparing testimony acceptance practices directly: Take for this purpose the f_is to be such testimony acceptance practices. Then f_o is a meta practice, a meta testimony acceptance practice which is reliability based. The optimality results of meta-induction show that such a meta testimony acceptance practice is long run optimal. Hence, there is a reliability based testimony acceptance practice which is epistemically justified. The investigation of (Schurz 2012b) shows that even if one restricts *accessibility* of so-called *local meta-induction* to the epistemic neighbourhood (i.e. a Moore-neighbourhood), still expert knowledge spreads.

So, we have outlined a very general approach to the problem of testimony. As mentioned already, the details are spelled out in section 10.3 when we consider the specific case of peers testifying a disagreement. However, before that we also want to consider other cases of testimony, most notably the case where an expert testifies to a novice. Such cases will be our concern in the next section.

9.4 Experts and Novices

We can differentiate different cases of testimony by help of a rough characterisation of the testifier and the recipient of the testimony in terms of *expert* and *novice*. Combining these categories, we can differentiate four relevant cases as in table 9.2.

testifier	*recipient of testimony*	*case*
novice	novice	peers
expert	expert	peers
novice	expert	redundant/wise crowd
expert	novice	authority

Table 9.2: Different cases of testimony

Now, how to incorporate testimony in case of peers will be discussed in detail in section 10.3. The case of a novice testifying to an expert seems to be in principle epistemically redundant, since the expert's opinion is epistemically better off—note that our preliminary notions of *expert* and *novice* include expertise with regards to all aspects, so, e.g., an expert has all the evidence also a novice has, etc. (e.g., we foresee from cases of *citizen science* here where huge data sets are created by laymen and studied by experts). However, even if the expert has all the evidence that also the novice has, there might be cases where an expert can deploy novice testimony, particularly if the novice is itself performing epistemically well, but just not as well as the expert: Such novice testimony might be deployed by an expert in order to make use of a so-called *wise crowd effect*, where a bulk of novices can outperform the expert opinion, once the novices are independent of each other and minimally competent. We will discuss such cases in chapter 12. In this section we are concerned with the last case, the case of an expert testifying to a novice.

The case of an expert testifying to a novice is a quite extensively discussed case of testimony. The traditional discussions of the expert/novice case centre around two problems: First, the problem of how a novice can *identify* an expert. And second, the problem of how exactly a novice should *incorporate the testimony* of an expert.

Let us come to the first problem. We can illustrate the identification problem well by reference to Plato's *Charmides*, where Socrates discusses with Charmides whether a novice can identify an expert and hence satisfy a precondition for justifyably taking in an expert's testimony (Charmides, Plato 1997, p.657):

> *Socrates:* "When another person claims to know something, [. . .] our friend [will not] be able to find out whether he knows what he says he knows or does not know it. But he will only know this much, it seems, that the man has some science;" (170d)
>
> *Charmides:* "Apparently so."
>
> *Socrates:* "So neither will he be able to distinguish the man who pretends to be a doctor, but is not, from the man who really is one, nor will he be able to make this distinction for any of the other experts." (170e)

It seems that a *novice* cannot distinguish between an *expert* and a *pseudo-expert*, for which reason one might be sceptic about the justification of testimony from experts. The argument is as follows:

1. Only experts can distinguish (*Dist*) *experts* (*e*) from *pseudo-experts* (*p*), but not *novices* (*n*).
   Schematically: $Dist_{ep}(x) \rightarrow x = e \;\&\; x \neq n$

2. Only if one can make such a *distinction*, than she is *justified* in relying on testimony from experts (real ore pretending ones).
   Schematically: $J(B_x(p|Test(p))) \rightarrow Dist_{ep}(x)$

3. Hence, novices are *not* justified in using testimony from experts.
   Schematically: $x = n \rightarrow \neg J(B_x(p|Test(p)))$

This is a very rough schema based on a rough distinction between experts and novices, and before one can evaluate the sceptical argument against testimony in case of an expert testifying to a novice, the notion of *expertise* has to be characterised. Basically, in science it is all about evidence and the inferences we draw from evidence. Hence, it is natural to expect from an expert in comparison to a novice, that the expert has more evidence and better inferential skills than the novice. As is almost always the case in describing a comparative notion by help of more than one parameter, in reality this relation is not that strict: So, typically we also count someone as an expert in comparison to a novice, in case the novice might have some evidence which the expert lacks. However, for simplicity reasons we assume (weak) "dominance" regarding both parameters here: *e* is an expert relative to n (n is a novice relative to e), iff e has all the evidence that n has, and e is inferentially better than n. In accordance with Goldman (2011b), we can explicate the latter via the conditions of truth indication:

$$Pr(p|Test_e(p)) > Pr(p|Test_n(p))$$

and truth tracking:

$$Pr(Test_e(p)|p) > Pr(Test_n(p)|p)$$

Note that if e is an expert with regards to n, and given the objective interpretation of truth indication and truth tracking from section 9.2, we have it that $rel_e > rel_n$, and hence n should incorporate e's testimony $Test_e(p)$ in order to approach rel_e. Now, the problem of Plato's *Charmides* is about the impossibility of n to figure out whether rel is high or not, in particular, whether $rel_e > rel_n$ or not or whether $rel_{e_1} > rel_{e_2}$, where e_1 is an expert and e_2 is a pseudo-expert with regards to n. The latter case is also relevant in so-called *novice/2-experts* cases, i.e. a case where $rel_{e_1} > n < rel_{e_2}$, but where the experts provide contradicting or disagreeing testimonies as, e.g.: $Test_{e_1}(p)$ and $Test_{e_2}(\neg p)$. Here "the *novice/2-experts* problem is whether a *layperson* can justifiably *choose* one putative *expert* as more credible or trustworthy than the other with respect to the question at hand, and what might be the epistemic basis for such a choice?" (Goldman 2011b, p.116).

Goldman (2011b) discusses the following five sources of evidence for n to assess $e = e_1$ or e_1 in comparison with e_2:

1. Arguments of e_1 and e_2 for or against p
2. Agreement of *further* experts: $Test_{e_1}(p)$ and $Test_{e_3}(p)$ vs. $Test_{e_2}(\neg p)$
3. Appraisals by *meta*-experts:
   $Test_{e_3}(Pr(p|Test_{e_1}(p)) > Pr(p|Test_{e_2}(p)))$
4. Interests and biases of e_1 or e_2
5. Past track records of e_1 and e_2

Strategy 1 applies, whenever e can make some of the arguments for her beliefs transparent to n or e_1 can make them better transparent to n than e_2 can do. In the best case, n receives a so-called *ostensible rebuttal defeater* against the testimony of one of the experts (see Goldman 2011b, p.218), i.e. an argument in the expertise of n which allows her to rule out one of e_1 or e_2. In general, one can expect that by gaining arguments from the experts n increases expertise (decreases relative novelty by diminishing her distance from e_1 or e_2) and hence should end up with higher rel. Note, e.g., that the programme of the *Vienna Circle* was not only programmatic regarding science, but society in general and based on the idea that experts should be able to explain their theories also to novices. In accordance with strategy 1 such claims can be found in “The Scientific Conception of the World. The Vienna Circle” (see Verein Ernst Mach 1996) and also very explicit in (Neurath 2006, p.400):

> “Ein Physiker muß die Forderung eines geistvollen Denkers grundsätzlich erfüllen können: ‘Jede streng wissenschaftliche Lehre muß man in ihren Grundzügen einem Droschkenkutscher in seiner Sprache verständlich machen können.’
>
> [In principle, a physicist must be able to fulfil the following demand: ‘Every scientific doctrine must be such that it can be made comprehensible to a cabman in his own language.’]”

Strategy 2 concerns the case of several coinciding testimonies which, as we have seen in section 9.2, allow also for an increase in *rel*. Strategy 3 helps once one has an accurate estimation of the reliability of the meta expert e_3: rel_{e_3}. One can think of such a meta expert, e.g., as an institution that provides *academic degrees, professional accreditations* that reflect e_1’s and e_2’s training and competence (see Goldman 2011b, p.220). Strategy 4 is directly related with rel_{e_1} or rel_{e_2}. As we also have discussed above, this strategy is often employed at court where prosecutors and lawyers bring biases of testifiers to the fore in order to undermine the jury’s estimation of the testifier’s reliabilities. Finally, 5 concerns directly the success based way discussed in section 9.3 and allows for optimisation regarding *rel*. These are all strategies that seem to overcome the problem of a novice n in *identifying* an expert e.

Let us briefly come to the second problem, the problem of how to *exactly incorporate testimony of experts*. Our arguments in the preceding sections show that incorporating testimony proportional to the degree of reliability of the testifier as suggested by (THume) allows for epistemic justification: In case of a prediction game with testifiers we were able to employ results of meta-induction to show that (internalist) rel_o is long run access optimal compared to (internalist) rel_i. And in case of single testimonies of single testifiers one might go along with Goldman’s argument of increased veracity on average given externalist rel_i. This seems to answer the question of incorporation of testimony in favour of (THume).

However, there is an argument put forward in the literature on *epistemic authority* which seems to undermine the reliability-proportionality approach to testimony. It is an argument in favour of the so-called *preemption thesis* which states that the belief or disbelief of an epistemic authority or expert e in a proposition should completely replace all reasons for or against the proposition of an epistemic subject submitted to the authority like a novice n. *Preemption* was prominently defended for the practical domain by Raz (1988) who considered it as a definitional feature of an *authority* to provide preemptive reasons. Zagzebski (2012), Keren (2014) and Constantin and Grundmann (2018) defended *preemption* for the epistemic realm. Roughly speaking, preemption demands a novice n to completely take over the epistemic attitude of an authority or expert e. If we assume that the testimony of the expert e is not just course grained $Test_e(p)$ or $Test_e(\neg p)$, but that e testifies her degree of belief $Pr_e(p) = r$, then preemption can be formulated as an update rule, given the expert's testimony $Pr_e(p) = r$ as evidence:

$$Pr_n(p) = Pr_u(p) = Pr_i(p|Pr_e(p) = r) = r = Pr_e(p) \qquad \text{(TPreemption)}$$

The main argument in favour of (TPreemption) is that if n ends up with a different degree of belief than e, then n is *watering down* the expert's judgement and is not guaranteed to approximate the reliability of e:

> "Suppose I decide [... to] first make up my own mind independently of the 'authority's' verdict, and then, in those cases in which my judgment differs from its, I will add a certain weight to the solution favoured by it, on the ground that it, the authority, knows better than I. [...] How will I fare under this procedure? [...] We can expect that in the cases in which I endorse the authority's judgment my rate of mistakes declines and equals that of the authority [i.e. rel_n approximates rel_e]. In the cases in which even now I contradict the authority's judgment the rate of my mistakes remains unchanged, i.e. greater than that of the authority. This shows that only by allowing the authority's judgment to pre-empt mine altogether will I succeed in improving my performance and bringing it to the level of the authority." (see Raz 1988, p.68)

Now, let us assume for simplicity reasons that truth/probability indication of the expert e as estimated by novice n is an adequate measure for reliability rel, i.e.: $rel_e = Pr_n(Pr(p) = r|Pr_e(p) = r)$, where $Pr(p)$ is an objective probability or chance of p. Then, if an expert's rel_e is high, then by Bayesian update on evidence $Pr_e(p) = r$ also $Pr_n(Pr(p) = r)$ is high, which collapses to $Pr_n(p) = r$. Hence, (TPreemption) can be considered a special case of (THume). However, note that given such an interpretation of rel, (TPreemption) is much stronger. It excludes, e.g., a reliability-proportional weighting of an expert's testimony as we needed in order to employ optimality. If we assume, e.g., that the novice n faces two experts e_1 and e_2 with $rel_{e_1} > rel_{e_2}$, then preemption seems to demand to preempt in favour of e_1 since this allows for approximation of rel_{e_1} by rel_n which is better than approximating rel_{e_2}. However, the meta-inductive testimony acceptance practice f_o we outlined in section 9.3 needs balancing between $Pr_{e_1}(p) = r_1$

and $Pr_{e_2}(p) = r_2$ in order to cash out long run access optimality. So, preemption seems to put forward a problem for f_o: On the one hand it seems clear that in case of expert testimonies the novice should preempt in favour of the best expert, because otherwise n might fall short of approximating *rel* of the best expert. On the other hand we know that f_o can employ optimality only if it mixes according to the expert's testimonies (online regression), because otherwise it is prone to suboptimality (online classification). So, the question is, is something wrong with our meta-inductive application to testimony?

As we will show now, everything is perfectly fine, because *preemption* concerns cases which are not covered by meta-induction. To see this, let us focus on the argument of Zagzebski (2012) in favour of (TPreemption): She refers to the problem of *probability matching* (see Zagzebski 2012, p.115). The problem of *probability matching* is as follows (see Vulkan 2000, sect.2): Given two mutually exclusive options r (right) and l (left) that are randomly distributed with fixed probabilities $Pr(r)$ and $Pr(l)$, what is the right strategy to make a decision for one of the options? As empirical studies show, humans tend to perform the so-called strategy of *probability matching*. According to this strategy, the frequency of one's decisions for an option should match the probability of the options. So, if, e.g., r shows up 75% of the time and l only 25% of the time ($Pr(r) = 0.75$ and $Pr(l) = 0.25$), then humans, when asked which option to choose, tend to opt for r also 75% of the time and for l 25% of the time (see Gallistel 1993, chpt.11). Non-human animals like rats act differently: They perform a *take-the-most-frequent strategy* that favours exclusively that option which has a higher probability. So, after some phase of learning the probabilities of the example above, they opt for r exclusively. What is the rationale of both strategies? It is easy to demonstrate that the expected utility of the *take-the-most-frequent strategy* is maximal (see Vulkan 2000, sect.2): If we calculate the expected utility for an agent n having credences Pr_n as follows:

$$Pr_n(r) \cdot Pr(r) \cdot u(r) + Pr_n(l) \cdot Pr(l) \cdot u(l)$$

If we furthermore assume that r and l are jointly exhaustive and mutually exclusive in the considered probability space; and if we finally assume that the utilities of r and l are equal ($u(r) = u(l)$), then it turns out that having $Pr_n(r) = 1.0$ in case $Pr(r) \geq Pr(l)$ maximises the expected utilities for n: Due to these assumptions the expected utility for n is proportional to:

$$Pr_n(r) \cdot (2 \cdot Pr(r) - 1) + 1 - Pr(r)$$

So, if $Pr(r) \geq Pr(l)$, then $Pr_n(r) \cdot [0.0, 1.0] + [0.0, 0.5]$ (the possible values under this assumption) is maximised by $Pr_n(r) = 1.0$. And if $Pr(r) < Pr(l)$, then $Pr_n(r) \cdot [-1.0, 0.0] + [0.5, 1.0]$ (the possible values under this assumption) is maximised by $Pr_n(r) = 0.0$. On average, the highest, i.e. the maximal expected, utility is gained by the *take-the-most-frequent strategy*: In the example above it will decide in 75% of the cases correctly. *Probability matching* is on average below (although, of course, in single instances it might perform better).

Now, note that the reasoning by help of maximising expected utilities amounts to applying a 0-1 loss as defined in definition 3.10 for determining *rel*. Note also that in his argument Raz (1988) spoke of "number of mistakes" in considering *rel*, so also he has a 0-1 loss in mind. However, as we have argued in section 3.4, online regression presupposes a convex loss function ℓ (see definition 3.30). This is the reason why the argument in favour of preemption (TPreemption) does not run against the meta-inductive version of (THume), but is about a different case of the problem of testimony by experts. This concerns the general case. Now, in case of a best expert (as in the example provided above) meta-induction converges to *imitate the best* (see, e.g., the one-favourite method in Schurz 2008b), and so also such a case (of online classification) is covered by meta-induction.

Chapter 10

Epistemic Peer Disagreement

In this chapter the problem of epistemic peer disagreement is characterised and the traditional approaches to this problem are explicated. Afterwards, a meta-inductive solution is introduced in detail which covers cases of peer disagreement and testimony in general. Finally, the meta-inductive approach is linked to arguments and objections present in the literature on peer disagreement, it is indicated how these impact the meta-inductive solution, and it is outlined how they can be overcome.

Roughly speaking, two agents have an epistemic *disagreement* with respect to some proposition, if one of them believes the proposition, whereas the other disbelieves it (or suspends judgement on it). It is a case of epistemic *peer* disagreement, once the two agents are epistemic peers. Since in science it is all about evidence and inferences, epistemic peers are characterised by having the same evidence and inferential skills.

Now, as we see it, the problem of epistemic peer disagreement concerns partly a generalisation of the problem of testimony, and partly a restriction of it. The generalisation results from considering several testifiers. Recall that in the case of testimony investigations often centre around two agents as, e.g., expert and laymen. In contrast to this, arguments in the peer disagreement debate very often crucially rely on considering a group of testifiers, peers. This brings us also to the restriction: The restriction of the problem of epistemic peer disagreement in comparison to testimony consists in considering mainly cases of *disagreement*, i.e. testimonies which deviate from one's own epistemic attitude towards a proposition, whereas in the debate of testimony matching of the attitudes is also of relevance. A general approach to the problem of testimony and peer disagreement considers cases with several epistemic agents and provides an answer of how to deal with any form of testimony, might it be that of a peer or not, might it be in disagreement with one's own beliefs or not.

There are several proposals to resolve the problem of epistemic peer disagreement which concentrate on the question of how to incorporate evidence of such a disagreement. The main positions in this field are the *equal weight view*, the *steadfast view*, and the *total evidence view* (see Frances and Matheson 2018, sect.5). In this chapter we present a new argument in favour of the equal weight view. As we will show, this view results from a general approach of forming epistemic attitudes based on testimonies in an optimal way. This general approach concerns all cases of testimony, and hence provides also an answer to the question of how to incorporate testimony in general. With respect to epistemic peer disagreement, our argument shows that one can strengthen the basis for equal weighting massively from reasoning via epistemic indifference to reasoning via optimality.

We will proceed as follows: First, we provide a general characterisation of *epistemic peer disagreement* in section 10.1. Then, in section 10.2, we present the formal framework of the classical approaches to epistemic peer disagreement in detail. Subsequently, in section 10.3, we expand the framework to the meta-inductive setting and provide our argument in favour of the equal weight view. There we also show how the other approaches fail to deal with the optimality argument. Finally, in section 10.4 we list some possible objections to our meta-inductive solution to the problem of peer disagreement and outline how they might be overcome.

10.1 The Problem of Epistemic Peer Disagreement

Two peers (more on the notion of *peers* below, for now we suppose that peers have the same evidence and inferential skills) have an epistemic disagreement regarding a proposition, if their epistemic attitudes towards the proposition differ. So, e.g., an agent a_1 might believe a proposition p whereas a peer a_2 disbelieves or suspends judgement regarding p. Or a_1's degrees of belief in p might be different from that of her peer a_2, etc. The question of how to deal with such a disagreement is the problem of epistemic peer disagreement.

Several proposals to resolve this problem have been put forward in the literature. Most of them mainly concentrate on the question of if, and if so, to what extent one should incorporate evidence of such a disagreement in forming an epistemic attitude towards a proposition. A classical position is the so-called *conciliatory view* which suggests to generally incorporate such evidence (see, e.g., Christensen 2007; Elga 2007; Feldman 2007). A position at the other end of the spectrum is the so-called *steadfast view* which suggests to generally not incorporate such evidence (see, e.g., Rosen 2001). In between are views that suggest to sometimes incorporate such evidence, and sometimes not or to incorporate such evidence from case to case differently (see, e.g., *the total evidence view* in T. Kelly 2011). The *intensive* debate of how to resolve epistemic disagreement lasts already for more than a decade now (it was initiated by Feldman 2007; clearly, systematic discussion of epistemic disagreement started much earlier; central is, e.g., Lehrer and Wagner 1981), and still there is little hope that *the disagreement among epistemologists* about epistemic peer disagreement will be resolved—at least regarding this matter *steadfasters* seem to have won, although there is also disagreement about this (see Elga 2010).

In this chapter we want to employ the framework of meta-induction and present a new argument in favour of the most prominent conciliatory view, namely the *equal weight view*. According to this view, one should assign the same weight to one's peer's epistemic attitude and to one's own. A prominent argument for equal weighting stems from a principle one might want to call the *principle of epistemic indifference*: If the epistemic attitudes of n individuals are, regarding their rational formation, epistemically indistinguishable (i.e. the individuals are epistemic peers), then each attitude should be assigned the same weight and, thus, $1/n$ if they are supposed to add up to 1. Similarly, as there is a big debate about such an indifference principle in statistics, there is also a big debate whether such a principle of indifference applies to the epistemic realm. *Steadfasters* typically deny this. What is missing is a principled argument why, in the case of peer disagreement, the principle of epistemic indifference should be applied. Why should an agent assign equal weight to both attitudes, her peer's and her own? We show in this chapter that the equal weight view results from consequentialist considerations: The equal weight view is an *optimal* strategy to resolve disagreement between peers.

Schurz (2012a) indicated already an application of his theory of meta-induction to the debate on *fundamental* disagreement, where epistemic agents "disagree in their underlying *cognitive system* [… i.e. they disagree on] fundamental principles of reasoning that determine the criteria for justification". He suggested to resolve such disagreements by help of applying methods that are "universal in the sense of being reasonable in every cognitive system" (Schurz 2012a, p.343 and p.346). As we show, this suggestion can be also expanded to the general case of epistemic peer disagreement and is even decisive regarding the single positions in the debate. We will argue for the claim that in the case of peer disagreement the underlying access-optimal meta-method is instantiated by the *equal weight view*, whereas the *steadfast view* as well as the *total evidence view* instantiate meta-methods that are not access-optimal. So, from an epistemic engineer's point of view, equal weighting has an important advantage over its competitors. In the next section we are going to embed the main approaches to peer disagreement into our setting of prediction games.

10.2 The Three Main Approaches

Epistemic peer disagreement with respect to a proposition p is that special case of disagreement, where two epistemic peers differ in their epistemic attitudes towards p. The explication of the *differentiæ specificæ* 'epistemic peer' and 'epistemic attitude' is crucial for our understanding of how to resolve the problem of epistemic peer disagreement. Let us briefly recall how they are understood in the debate:

> "Let's say that people are epistemic peers when they are roughly equal with respect to intelligence, reasoning powers, background information, etc." (Feldman 2007, see p.201)

> "[Peer disagreement is at hand if] each of you has access to the same [...] statistics, [...] reports, and so on, and has no other relevant evidence. Furthermore, you count your friend as an epistemic peer—as being as good as you at evaluating such claims." (see Elga 2007, p.484)

> "[If I] suppose that my friend and I have had long discussions in which we share every bit of evidence we can think of that's relevant [...] and suppose further that I have good reason to believe that my friend and I are equally intelligent and rational, and that I know of no general reason [...] to think either of us is especially likely to be particularly good, or bad, at reacting to evidence[, ... then] my friend seems to be what some have called an 'epistemic peer'." (Christensen 2007, pp.188f)

That two individuals, a_1 and a_2, are epistemic peers is best captured by the slogan: *The same evidence and inferential skills are what make epistemic peers*. Clearly, traditional principles of rationality in epistemology and philosophy of science cover cases where there is some unbalancing in one or both of these factors. Here is a rough line of argumentation one could follow, but, clearly, there are many other consensual ways of justifying resolutions of disagreement cases that are disagreements among non-peers like expert and layman: If a_1 has more relevant evidence than a_2 while, generally, a_1 is as good as a_2 in making (statistical) inferences, then the statistical *principle of total evidence* favours a_1's epistemic attitude over that of a_2 (see Carnap 1947, p.141). If the evidence of a_1 is different from that of a_2 without one of them being contained in the other, then principles for merging statistics with different "reference classes" might apply (see Kyburg (Jr.) and Teng 2001). If a_1's inferences are more reliable than that of a_2, while at the same time a_1's evidence is also at least as comprehensive as that of a_2, then reliabilistic principles favour the epistemic attitude of a_1 over that of a_2 (see

Goldman 2014). And, finally, if a_1's inferences are more reliable than that of a_2, while a_2 has more relevant evidence, then reliabilistic principles and the principle of total evidence suggest to form an epistemic attitude based on a_2's evidence by help of a_1's method of inference.

The same traditional principles of rationality, however, leave it quite open of how to cope with cases where both factors, evidence and inferential competence, are equally balanced among the individuals. There are approaches that seem to rule out the possibility of such cases of disagreement. So, e.g., Carnap's *theory of logical probability* as briefly discussed in section 5.2 suggests that there is only one rational epistemic attitude towards a hypothesis given some evidence (see Carnap 1950/1962, note that strictly speaking also Carnap allows for a continuum of inductive methods with a parameter for the speed of learning which does not satisfy *uniqueness*). However, this proposal is quite non-consensual as are many other proposals within the so-called *uniqueness* framework of epistemic justification. Especially if one considers, e.g., fine-grained epistemic attitudes and the widespread Bayesian framework, there seems to be little hope of vindicating a general uniqueness principle according to which "a body of evidence justifies at most one proposition out of a competing set of propositions" (see Feldman 2007, p.205). In the Bayesian case the relaxed treatment of *priors* allows for rationalising many different epistemic attitudes. Nevertheless, also within this paradigm of relatively casual justification authors aim at strictly excluding the possibility of "*anything goes!*". A case in point are investigations of convergence results that aim at predicting a phenomenon called *washing out of priors*: "Although your [prior] opinion about future behaviour of a coin may differ radically from your [peer]'s, your opinion and his will ordinarily be transformed by application of Bayes' theorem to the results of a long sequence of experimental flips as to become nearly indistinguishable" (see (Edwards, Lindman, and Savage 1963), cited according to Earman 1992, p.141). So, the idea of such convergence results is to show that under certain conditions as, e.g., the condition that all peers incorporate the same evidence and also in the same way—namely by strict conditionalisation—, the more evidence is accumulated, the more the initially disagreeing positions converge to each other (another important condition is that of the peers being equally dogmatic—see Earman 1992, chpt.6). Hence, although Bayesian orthodoxy denies a uniqueness thesis as, e.g., held by Feldman (2007), Bayesian epistemologists stressing the usefulness of convergence results seem to agree with a principle one might call *long run* or *limiting uniqueness*: The more evidence two epistemic peers gather, the less diverse and more similar their rational epistemic attitudes are. Trivially, in the limiting case where all the evidence is on the table, there is only one single epistemic attitude towards a hypothesis about the evidence rationalised. Our optimality argument in favour of the equal weight view will be also such a *long run limiting case* argument.

Before we come to our argument, we need to say a little bit more about the second ingredient of a case of peer disagreement as well as the different approaches to resolve it. But step by step! Epistemic attitudes might be considered on a nominal scale as, e.g., the traditional all-or-nothing belief, disbelief, and suspension of judgement; on an ordinal scale as, e.g., the epistemic rank κ of one belief in comparison to another (see, e.g., Spohn 2012); or on a cardinal scale as, e.g., the degree of belief Pr. In (T. Kelly 2011) it is convincingly argued that, in order to reasonably compare different approaches to epistemic peer disagreement including a "weighting" approach, one needs to consider epistemic attitudes on the cardinal scale: E.g., if peer a_1 disbeliefs proposition p and peer a_2 suspends judgement regarding p, what would be an intermediate position (see T. Kelly 2011, sect.2)? Similarly for an ordinal ranking: If a_1's belief in p has rank n, and a_2's belief in p has the immediately following rank $n+1$, what would be an intermediate position here? One might even guarantee that in case of disagreement among two peers there is always an intermediate position available. This is achieved, e.g., be considering only cases of so-called *strong disagreement*, where one agent believes p whereas her peer disbelieves p. However, also given such a guarantee one can easily construct multi agent cases where an intermediate position is missing (see T. Kelly 2011, p.188): Consider the case of a_1 believing p, and by this strongly disagreeing with a_2 and a_3 who both disbelieve p. How can such attitudes be "weighted" adequately? If a_1 is supposed to resolve the disagreement by also disbelieving p, then her position seems to be *under*-weighted. If she resolves towards believing p or suspending judgement on whether to believe p or not to believe p, her position seems to be *over*-weighted. In order to formulate all classical responses to the problem of epistemic peer disagreement adequately, it seems necessary to consider the problem on a cardinal scale. So, the epistemic attitudes under investigation are degrees of belief in a proposition. Disagreement regarding these attitudes consists of two individuals having different degrees of belief in a proposition.

Finally, let us come to the classical responses! Following T. Kelly (2011) we want to distinguish two types of evidence, namely *first order evidence* and *higher order evidence*. As higher order evidence, sometimes also called *psychological evidence*, might serve any kind of information which is about the degrees of belief of an epistemic agent. So, e.g., in case of peer disagreement it might be the proposition that agent a_1's degree of belief in a proposition p ($Pr_1(p)$) is such and such. All other information not about degrees of belief of an epistemic agent might qualify as first order evidence, sometimes also called *non-psychological evidence* (see T. Kelly 2011, p.194).

According to a slogan of Feldman (2007, p.208), "evidence of evidence is evidence". More specifically, Richard Feldman thinks that "evidence that there is evidence for P is evidence for P." As is shown by Fitelson (2012), this principle runs against the common probabilistic notion of evidential support in the sense of probabilistic increase. However, if we consider only a weak implication of this slogan, namely that higher order evidence also counts as evidence, then in principle all approaches to the problem of epistemic peer disagreement can agree on this. They just differ along the line of how to incorporate this kind of evidence. If we spell out 'incorporate', as is often suggested, in terms of linear weighting (see Elga 2007), then we can describe the case of peer disagreement as one of finding a correct weighted update of one's degrees of belief given some higher order evidence. Another important way of weighted updating consists in geometrical weighting (see, e.g., Dietrich and List 2016). In contrast to linear weighting, this method satisfies, e.g., the desideratum of commutativity with learning—which, sloppily speaking, means in this context that one can commute the order of updating on first order evidence and on higher order evidence without any difference in the result (see, e.g., Brössel and Eder 2014, p. 2368). A geometrical method which satisfies this constraint and was introduced recently is, e.g., *Upco*: *updating on the credences of others* (Easwaran et al. 2016). However, since the three classical approaches to epistemic peer disagreement can be described on the basis of linear weighting, we will not consider these ways of incorporating higher order evidence in this chapter. We can model the case of epistemic peer disagreement as follows:

- Let $Pr_1^0, \ldots, Pr_n^0$ be the prior probability distributions of n agents, i.e. their degrees of belief regarding all propositions of an algebra before they receive any evidence.

- And let us assume that $Pr_i^1, Pr_i^2, Pr_i^3, \ldots$ $(1 \leq i \leq n)$ are the agents' posterior probability distributions after updating on first order evidence $e^1, e^2, e^3, \ldots$. (The superscript indices mark the rounds of update; the prior round without any evidence has the index 0).

- Since in the case of epistemic peer disagreement it is assumed that the peers share all their evidence, we assume that their sequences of updating are synchronous, i.e. a_i updates on evidence e at the same round as a_j updates on e (for all $1 \leq i, j \leq n$, $e \in \{e^1, e^2, e^3, \ldots\}$).

- Now, in case of epistemic peer disagreement between a_i and a_j regarding some proposition p it holds at some round $t \in \mathbb{N}$: $Pr_i^t(p) = r_i \neq r_j = Pr_j^t(p)$.

- But not only this: Furthermore, the agents get aware of this—they receive in the same round (but in a second phase) higher order evidence $e_h^t = (Pr_1^t(p) = r_1 \& \cdots \& Pr_n^t(p) = r_n)$ and have to update on it. Thus, in this setting the problem of epistemic peer disagreement can be formulated as follows:

$$Pr_i^t(p|e_h^t \& e^t) = Pr_i^t(p|Pr_1^t(p|e^t) = r_1 \& \cdots \& Pr_n^t(p|e^t) = r_n) =?$$

- As we have stated above, incorporation of such higher order evidence is often described as a form of linear weighting. So it holds:

> **Epistemic Peer Disagreement**
> There are $w_1^t, \dots, w_n^t$ such that:
>
> $$1.\ Pr_i^t(p|e_h^t \& e^t) = \sum_{j=1}^{n} w_j^t \cdot Pr_j^t(p|e^t) \qquad \text{(EPD)}$$

This is our model of the problem of epistemic peer disagreement. Note that in this model the weights are round-dependent as they can vary among the rounds. We will see later on that this is crucial for motivating our optimality-argument. In most investigations the weights are considered to be constant, i.e. round-independent. However, as we will see later on, this expresses the model assumption that the factor of inferential competence does not change. We want to be flexible on this. The more since we can "soft-code" this assumption by simply demanding $w_i^{t+1} = w_i^t$ (for all $0 < t \in \mathbb{N}$ and $1 \leq i \leq n$). Note also that the problem of epistemic peer disagreement is not intended for covering round 0, i.e. the round prior grasping any evidence. This is necessary, because otherwise, e.g., according to the equal weight view epistemic peers would always need to start from the same prior distributions. An assumption too strong to be generally subscribed to this view. Finally, observe that the assumption of shared evidence is "hard-coded" in the model: All agents update at the same round on the same evidence. This is due to the fact that in our discussion we need no flexibility regarding differences in evidence.

Given this framework, we can describe three classical approaches to the problem of epistemic peer disagreement with the help of the following specifications: The equal weight view claims that in case of peer disagreement the epistemic attitudes of all peers should get equal weight. So, the weights of the model above are specified as follows:

> **Equal Weight View**
> Among peers the weights are equal: 1. of (EPD) and:
>
> $$2.\ w_1^t = \dots = w_n^t = 1/n \ \text{ (for all } 0 < t \in \mathbb{N}) \qquad \text{(EWV)}$$

Note that from this it follows that, according to (EWV), the weights among peers remain constant in all cases: $w_i^{t+1} = w_i^t$ (for all $0 < t \in \mathbb{N}$ and $1 \leq i \leq n$). Considering the impact of first order evidence and higher order evidence in an agent's forming of an epistemic attitude, it is easy to see that higher order evidence can swamp first order evidence: Once an agent becomes aware that she is in a situation of epistemic peer disagreement, her update is determined by the higher order evidence alone: She computes her new credence simply as the average of $Pr_1^t(p|e^t) = r_1, \dots, Pr_n^t(p|e^t) =$

r_n. There is no need for her to recall how she ended up with $Pr_i^t(p|e^t) = r_i$ before (which was by conditionalising on first order evidence). In this sense, according to (EWV), what matters is only higher order evidence. The most prominent proponents of (EWV) are Christensen (2007) and Elga (2007). A coarse-grained version of the view for a nominal scale is held by Feldman (2007).

The steadfast view points in the completely opposite direction: According to it, higher order evidence plays no role in forming one's epistemic attitudes. Technically, this is implemented by simply weighting one's own epistemic attitude fully, whereas that of one's peers get no weight:

> **Remain Steadfast View**
> Among peers one's own position gets full weight: 1. of (EPD) and
>
> $$\begin{aligned} &2.\ w_i^t = 1 \text{ and } w_j^t = 0 \\ &\quad(\text{for all } 0 < t \in \mathbb{N}, j \in \{1, \ldots, n\} \setminus \{i\}) \end{aligned} \tag{RSV}$$

Note that as above, these weights among peers remain constant. Since updating on higher order evidence does not change anything in the degrees of belief of an agent, according to (RSV) what matters is only first order evidence in forming one's epistemic attitudes. Perhaps the most prominent proponent of (RSV) is Rosen (2001).

Finally, we also want to model the total evidence view: According to this view, only taking into account either higher order evidence or first order evidence provides no adequate response to the total evidence available:

> "The equal weight view and the no independent weight view [i.e. the steadfast view] both suffer from the same fault: they embody overly simple models of how one's first order evidence and one's higher order evidence interact in determining facts about what it is reasonable to believe all things considered. [. . .] Rather, what it is reasonable to believe depends on both the original, first order evidence as well as on the higher order evidence that is afforded by the fact that one's peers believe as they do." (see T. Kelly 2011, p.201)

Now, it is clear that this description of the total evidence view does not automatically ask for some *linear* "interaction" between first and higher order evidence. However, the arguments and further phrasing of it (see, e.g., *swamping*, *insubstantial and substantial evidence*, *equally strong pieces of evidence*, *greater proportion of our total evidence* etc. in T. Kelly 2011, pp.201ff) seem to grant such a *linear* "interaction". This the more, as we will see now, as the weights can be varied from case to case:

Total Evidence View
Among peers one's own position might be partly or fully influenced by the other peers' positions: 1. of (EPD) and

$$\begin{aligned} 2.\ & w_1^t + \cdots + w_n^t = 1, \\ & w_j^t \geq 0 \ \ (1 \leq j \leq n) \end{aligned} \tag{TEV}$$

Note that (TEV) just guarantees that the weights used in resolving epistemic peer disagreement allow for linear weighting. These weights among peers need not be constant for different rounds, i.e., different situations of disagreement. And they allow for some impact of one's higher order as well as one's first order evidence.

Considering the ways of determining the impact of higher order evidence in case of an epistemic peer disagreement, these are the three available possibilities as illustrated in figure 10.1: Higher order evidence is not at all taken into account (RSV), almost only higher order evidence is taken into account (EWV) or higher order evidence is variably taken into account (TEV).

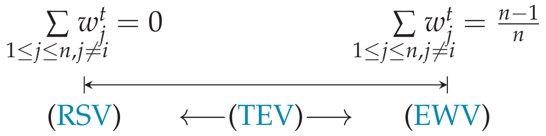
$\sum_{1 \leq j \leq n, j \neq i} w_j^t = 0$
$\sum_{1 \leq j \leq n, j \neq i} w_j^t = \frac{n-1}{n}$
(RSV)
⟵(TEV)⟶
(EWV)

Figure 10.1: Spectrum of possible ways of incorporating higher order evidence in case of epistemic peer disagreement: Let i by the agent updating her degrees of belief in case of such a disagreement and let $n-1$ be the number of her peers. According to the *remain steadfast view*, the weights used for updating based on higher order evidence is 0. According to the *equal weight view*, given a high number of (independent) peers, i's first order evidence (weighted by $1/n$) can be swamped by the higher order disagreement evidence (weighted by $(n-1)/n$). The *total evidence view* allows for any linear weighting of higher order evidence within this spectrum. Note that in principle the total evidence view could also extend the spectrum to the right by *boosting* peers via higher weights ($w_j^t > w_i^t$). However, since in the traditional debate no argument covers this case, we also do not consider it here.

The arguments for and against each of these views are well discussed in the debate on epistemic peer disagreement. We will not present and discuss them here in detail. Rather, given our formal model, we want to add a further argument to the debate which strikes us as decisive with respect to (EWV). Afterwards, we will recap some arguments against (EWV) when we discuss some possible objections to our argument.

10.3 The Meta-Inductive Approach

As mentioned earlier, one argument for the equal weight view (EWV) originates from indifference-considerations: If the epistemic attitudes of some peers are indistinguishable with respect to their underlying evidence as well as their inferential skills applied to the evidence, why should there be a difference with respect to their epistemic impact in updating one's degrees of belief, once one becomes aware of a disagreement? The assumption, that all agents share the same evidence, is already *hard-wired* in our model: At each round all agents update on the same evidence. But how can we express that the peers have the same inferential skills? We want to suggest to implement this into the model by help of a reliabilistic measure of "inferential" or predictive success.

The idea is as follows: Each agent has to make a prediction about the truth value of a proposition p^t or a set of propositions $p^t_1, \ldots, p^t_k$ at each round t. These predictions $Pr^t_i(p^t)$ or $Pr^t_i(p^t_1), \ldots, Pr^t_i(p^t_k)$ are based on the shared evidence e^t as well as the individual inference or prediction method of agent Pr_i. We assume that afterwards all predictions are revealed to all agents and might serve as higher order evidence e^t_h for the same round t. Each agent has to make again a prediction about p^t or $p^t_1, \ldots, p^t_k$ respectively. Now, we assume that at the end of a round t the truth value of p^t or $p^t_1, \ldots, p^t_k$ is settled—and for simplicity of expression we assume also that it is revealed to all agents. So, the cases to which the argument presented here applies are no cases of *deep disagreement*. Rather, they are cases in which further evidence can dissolve the disagreement. Later on we will discuss ways of relaxing the assumption that the outcomes are revealed to the agents.

In this dynamics each round t consists of three phases: a phase of updating on first order evidence e^t, a phase of updating on higher order evidence e^t_h, and a phase where the truth value(s) are revealed. Figure 10.2 illustrates how the specified model of making inferences and predictions looks like. We consider here only cases of predicting one proposition in each round (p^t). Later on we hint at generalising the model to cases where one makes predictions about several related propositions or a whole algebra in one and the same round ($p^t_1, \ldots, p^t_k$).

Now, as we outlined above, given the truth values of the propositions in question one can define a measure for the reliability of an agent's predictions and inferences by measuring the average closeness of the agent's predictions and inferences to the truth. For this purpose we take one minus the squared difference of both values, the inferred one and the truth value of the proposition p in question, i.e. $val(p)$, where truth is represented by $val(p) = 1$ and falsity by $val(p) = 0$. Our choice of a quadratic scoring measure (see Brier 1950) is due to its favourable theoretical properties like being a *so-called* proper scoring measure (at the end of this chapter we indi-

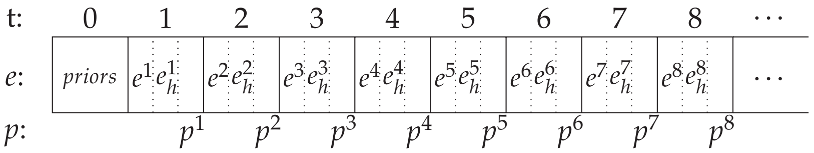

Figure 10.2: Model of peer disagreement: At each round all agents receive first order evidence e^t and have to make their inferences based on this evidence regarding propositions p^t. Afterwards, all agents receive information about the other agents' inferences, i.e. higher order evidence e^t_h. They have to make again inferences about the propositions in question which they might base on this further evidence. At the end of each round the truth value of the proposition(s) in question is revealed.

cate how this approach to epistemic peer disagreement can be generalised to probabilistic cases and proper scoring rules are particularly suited for scoring probabilistic predictions). As we have seen in part I of this book, the optimality result we apply here does not depend on this choice. Rather, it holds for all scoring measures which are based on an underlying *convex* loss function. We define the reliability or success s^t_i of an inference or prediction method of agent i up to round t regarding proposition(s) p as follows:

$$s^t_i = \frac{\sum\limits_{0<u\leq t} 1 - (val(p^u) - Pr^u_i(p^u|e^t))^2}{t} \tag{10.1}$$

For illustrative purposes, consider the outcomes, inferences, and reliabilities according to table 10.1: An agent who gets all inferences absolutely right has a reliability of 1, whereas an agent who gets all inferences absolutely wrong has a reliability of 0. All other kinds of inferences are strictly in-between this interval.

Now, let us explicate the notion of an *epistemic peer* in this model. As we discussed in the preceding section, two agents are epistemic peers iff they possess the same (i.e. shared) evidence and equal inference skills regarding the evidence. Since evidence sharing is hard-wired in the model, all agents of the model are peers in this respect. But how about the other relevant attribute? It seems plausible to assume that equal inferential skills can be expressed by equal reliabilities: According to this model, two agents a_1 and a_2 are equally skilled regarding inferences on shared evidence at round t, if their reliabilities $s^t_{a_1}$ and $s^t_{a_2}$ match, i.e.: $s^t_{a_1} = s^t_{a_2}$. So, the question of how to update one's degrees of belief on higher order evidence about one's epistemic peers' epistemic attitudes results in the question of how to update, given $s^t_i = s^t_j$ (for all $1 \leq i, j \leq n$)? Now, we can make this assumption explicit in our model by restricting condition 1. of (EPD) to cases of disagreement with peers that have the same reliabilities:

$t:$	0	1			2			3			4		
phase :		1	2	3	1	2	3	1	2	3	1	2	3
p^t :				0			0			1			1
Pr_1^t		0.5	0.5		0.25	0.25		0.125	0.125		0.25	0.25	
Pr_2^t		0.5	0.5		0.5	0.375		0.375	0.25		0.25	0.25	
Pr_3^t		0.0	0.0		0.0	0.0		1.0	1.0		1.0	1.0	
Pr_4^t		1.0	1.0		1.0	1.0		0.0	0.0		0.0	0.0	
s_1^t				0.75			0.844			0.641			0.590
s_2^t				0.75			0.805			0.682			0.621
s_3^t				1.0			1.0			1.0			1.0
s_4^t				0.0			0.0			0.0			0.0

Table 10.1: Example of applying the reliability measure: Pr_3 is always right and hence receives a degree of reliability of 1. Pr_4 is always wrong and hence has reliability 0. Pr_1 is an inductive method not updating on higher order evidence (inferences in phase 1 equal that of phase 2). Pr_2 is also an inductive method but updates on higher order evidence (inferences in phase 1 differ from that in phase 2).

Epistemic Peer Disagreement (further specified)
There are $w_1^t, \ldots, w_n^t$ such that:

$$1^*.\ Pr_i^t(p|e_h^t \& e^t) = \sum_{j=1}^{n} w_j^t \cdot Pr_j^t(p|e^t) \qquad \text{(EPD}^*\text{)}$$
$$\text{if } s_1^t = \cdots = s_n^t.$$

We want to highlight that we intend to interpret the reliabilities s_i primarily in an operational way, i.e. as being empirically accessible. So, the question of whether two agents are peers or not can be empirically addressed. However, we do not want to exclude a more normative interpretation of our model: One might, e.g., also argue that the notion of *peerhood* is normative insofar as it is about something we put forward or assume without empirical backup; in this case one might interpret the reliabilities s_i not empirically: That $s_i^t = s_j^t$ is to be interpreted as the *normative assumption* that agents i and j are equally reliable.

Note that the problem of epistemic peer disagreement as defined above is relevant only for cases where the reliabilities s_i^t of the peers match. Note also that in this framing of the problem the connection between the weights and these reliabilities is completely unconstrained—this is what allows us to spell out the different approaches to peer disagreement by simply defining different constraints on these weights. Furthermore, higher order evidence e_h^t consists no longer only in knowing that one's peers disagree with respect to one and the same first order evidence, rather one also knows that one's peers are equally reliable ($e_h^t = (Pr_1^t(p) = r_1 \& \cdots \& Pr_n^t(p) = r_n \& s_1^t = \cdots = s_n^t)$). In the following we assume that all considered approaches to epistemic peer disagreement, i.e. (EWV), (RSV), and (TEV), are based on (EPD*). We want to show now that the equal weight view (EWV) is a specific instance of a general rule on incorporating higher order evidence which is proven to be optimal. We also want to show that the remain steadfast view (RSV) as well as the total evidence view (TEV)—in case it deviates from (EWV)—are, in terms of reliability, epistemically suboptimal. Here are the details: The reliability measure as defined in equation (10.1) can be considered as measuring the epistemic performance of an agent. As we indicated in table 10.1, the best performance possible up to round t is given if an agent i's reliability s_i^t is 1. This means that all of her inferences were correct up to t. If we take the set of all relevant epistemic propositions to be countable infinite, then we can interpret the best possible performance simply as that of an agent i having reliability $s_i^{t\to\infty} = 1$. As we have argued in part II of this book, even an ideal agent might miss the target of achieving the best possible performance. So, $s_i^{t\to\infty} = 1$, i.e. absolute learnability, is not a requirement we usually put forward for rational agents. However, a constraint which is often put forward results from comparative considerations on rationality: As in many other domains, the term 'rationality' can be also considered as expressing efficiency. So, e.g., in decision theory, a decision is rational if it is shown to maximise expected utility. I.e.: given a set of decisions $d_1, \ldots, d_k$, choosing a d_i is rational, if the expected utility eu of d_i, i.e. $eu(d_i)$, is maximal: $eu(d_i) \geq eu(d_j)$ $(1 \leq j \leq k)$. Similarly to the classical realm, we can demand for the social epistemological realm relative learnability, i.e. *long run optimality*: An inference method Pr_i is epistemically rational, if its reliability is *long run optimal* compared to all available inference methods $Pr_1, \ldots, Pr_n$ in the sense that $s_i^{t\to\infty} \geq s_j^{t\to\infty}$ $(1 \leq j \leq n)$. Given this epistemic constraint, one can show that (EWV) satisfies it, whereas (RSV) as well as (TEV) fail to do so. For the case of (EWV) we will show this by employing the general optimality results of the theory of meta-induction, as described in part I and applied in part II of this book. For showing the suboptimality of (RSV) and (TEV) we will provide examples where both of them fail to produce optimal inferences.

Let us start with the optimality of (EWV)! It is not hard to see that our model of epistemic peer disagreement is an instance of the more general model of the dynamics of epistemic inferences and predictions from the preceding parts. Here is, how we can employ the optimality results from before: Let us think of the probability functions $Pr_1, \ldots, Pr_n$ as object inference methods in the sense that whenever they conditionalise on first order evidence, their inferences are functionally independent, i.e. the definition of such an inference method Pr_i based on first order evidence contains no reference to one of the other $Pr_1, \ldots, Pr_n$. Now, as described above (equation (10.1)), for each method we can define a reliability measure s_i^t. Based on this reliability measure, we can define a meta-method Pr_m whose inferences or predictions are weighted averages of the object-methods' inferences and predictions. For technical reasons (see the short description in figure 3.6), we have to cut off those object-methods which performed worse than the meta-method in the past. So, the considered reliabilities for the meta-method Pr_m are defined as follows:

$$s'_{i,t} = \begin{cases} s_i^t, & \text{if } s_i^t \geq s_m^t \text{ or if for all } 1 \leq j \leq n\text{: } s_j^t \leq s_m^t. \\ 0, & \text{otherwise.} \end{cases} \tag{10.2}$$

Here s_m^t is the reliability of the meta-method Pr_m up to round t. By normalising the considered reliabilities, we get weights for the meta-method Pr_m:

$$w_i^t = \frac{s'_{i,t-1}}{\sum\limits_{j=1}^{n} s'_{j,t-1}} \tag{10.3}$$

Finally, by linear weighting we can recursively define the meta-method Pr_m as follows:

$$Pr_m^t(p) = \sum_{i=1}^{n} w_i^t \cdot Pr_i^t(p) \tag{10.4}$$

The inference or prediction of Pr_m at round 0 might be arbitrarily chosen—it might be, e.g., the unweighted average of the object-level inferences.

Now, note that Pr_m is an attractivity weighting meta-inductivist (f_{ami}), so we can simply apply the optimality result of this method and get (by theorem 3.40):

Theorem 10.1.

$$\lim_{t \to \infty} s_m^t - max(s_1^t, \ldots, s_n^t) \; \geq \; 0$$

So, by this theorem we know that Pr_m is an inference method that performs optimal in the long run compared to all available inference methods $Pr_1, \ldots, Pr_n$. Given our epistemic constraint, this provides a reason for considering Pr_m to be a rational inference method, i.e. to be epistemically justified.

Now, it is easy to see that Pr_m is an inference method based only on higher order evidence. Furthermore, there are no constraints for update on evidence or the reliabilities, so Pr_m is a long run optimal method for all cases, cases of no disagreement, cases of disagreement, cases of disagreement among epistemic peers, and cases of disagreement among non-peers.

What is important to note is that in the specific case of epistemic peer disagreement, Pr_m coincides with (EWV): Given $s_1^t = \cdots = s_n^t$ as mentioned in condition 1^* of (EPD*), it follows from the definition of Pr_m via equations (10.2) to (10.4) that $w_1^t = \cdots = w_n^t = 1/n$. Hence, (EWV) instantiates Pr_m for the specific case of epistemic peer disagreement. Since Pr_m is shown to be optimal regarding all cases of agreement and disagreement, (EWV) is optimal regarding cases of epistemic peer disagreement. Hence, according to our epistemic constraint, (EWV) is epistemically justified regarding cases of epistemic peer disagreement.

Let us come to a suboptimality proof of the alternative approaches to epistemic peer disagreement: (RSV) and (TEV). For this purpose it suffices to provide examples where these approaches fail to be optimal in the long run. It is easy to construct an "environment" which favours the competitors of (RSV) and (TEV), although both methods might, from time to time, catch up in reliabilistic terms. Consider, e.g. a setting with two agents, having inference methods Pr_1 and Pr_2, where the latter represents a *steadfaster* (RSV) or an agent considering the total evidence (TEV) which does not coincide with (EWV). Now, let us assume that out of three pairs of inference rounds, one is a round with epistemic peer disagreement and the other two are rounds with disagreement among non-peers. We can easily think of inferences with reliabilities distributed as shown in table 10.2. But then, just by averaging the reliabilities, one can see that Pr_1 outperforms Pr_2, since in the long run (on average) it holds $s_1^{t\to\infty} = 0.50 > 0.497 = s_2^{t\to\infty}$. Note that the example is constructed in such a way that suboptimality results from not equally weighting in case of a peer disagreement: Agent 2 outperforms agent 1 before agent 1 becomes a peer and agent 2 performs suboptimally when deciding to not equally weight higher order evidence of agent 1's epistemic attitude.

t: (for any $0 < v \in \mathbb{N}$)	$2 \cdot v$	$2 \cdot v + 2$	$2 \cdot v + 4$	$\cdots$
s_1^t	0.50	0.50	0.50	$\cdots$
s_2^t	0.51	0.50	0.48	$\cdots$

Table 10.2: Example of the suboptimality of (RSV) and (TEV) due to not weighting equally among one's epistemic peers in case of epistemic peer disagreement: Pr_1 gets the inferences in 50% of the cases right, whereas Pr_2 is sometimes slightly better, then Pr_1 catches up and then, in the case of a peer disagreement, strategy Pr_2 of remaining steadfast or incorporating total evidence looses.

The optimality result for (EWV) shows that such a case cannot appear if one performs equal weighting. Clearly, an agent performing (EWV) can also be suboptimal compared to her competitors. The simplest case one might think of is a setting where no peer disagreement shows up, because the agents' reliabilities never match. However, this suboptimality of (EWV) is due to her being suboptimal already in cases of non epistemic peer disagreement. Regarding cases of epistemic peer disagreement—which are the cases the debate is mainly about—(EWV) is guaranteed to be optimal.

We want to conclude this section with a short note on generalising this result: In our model we presupposed that in each inference round t there is a proposition p^t the inference is about and whose truth value is revealed in the third phase of the round. The inferences of the agents might be considered to be probabilistic statements about p^t, since $Pr_i(p^t) \in [0,1]$. Now, this setting can be expanded also to probabilistic statements about not only one (or two) proposition(s) p^t (and $\neg p^t$), but also to a set of mutually exclusive and jointly exhaustive propositions $p_1^t, \ldots, p_k^t$. In this case the individual inferences concern probabilistic statements of the form $Pr_i(p_1^t) \in [0,1]$ and $\ldots$ and $Pr_i(p_k^t) \in [0,1]$, where $Pr_i(p_1^t) + \cdots + Pr_i(p_k^t) = 1$. So, each agent provides in each round a probability distribution. We will see in chapter 11 that also for such a setting a meta-method can be defined which proves to be optimal. A simple way of doing so is by considering as relevant for the reliability of an agent in each round t only that event of $p_1^t, \ldots, p_k^t$ which turned out to be true. Then weights are constructed out of these reliabilities and the resulting view for epistemic peer disagreement is again an equal weight view.

10.4 Objections, Replies, and Restrictions

Now we are going to embed our argument in favour of the equal weight view (EWV) a little bit more into the traditional debate about how to resolve epistemic peer disagreement.

First, let us discuss one objection one might want to put forward against our model: In order to measure the reliability of the epistemic agents, our model presupposes that the truth values of the propositions are revealed at some point in time to the epistemic agents. However, in case of *deep disagreement* there might be no possibility to get to know these values. So, how can our model be employed in these cases? Let us mention that a conditional interpretation of our model might be satisfying for resolving cases of deep disagreement too. The modification is as follows: At phase 3 the agents no longer receive the truth values of the propositions in question, but remain with their own estimations. In this case, the argument reads as follows: *if* the agents get their estimations of the reliability of the other agents in the setting right, *then* (EWV) proves to be *the* optimal strategy and by this is justified. Clearly, whether (EWV) is de facto long run optimal remains an open question. However, it seems that following this possible route to *optimality* is the best one can do, the more as even in case of deep disagreement it is supposed that *supposedly* peers have no disagreement about the status of their peerhood. So, even in case of deep disagreement *steadfasters*, *total evidentialists*, as well as *equal weighters* agree on the conditional part, namely that they got the estimations of the reliability right. Things get much trickier, of course, if there is even a disagreement about whether there is disagreement and if so, whether it is among peers. The traditional debate about epistemic peer disagreement has little to say about this and it seems that this is not without reason.

A general objection to (EWV) is provided by the so-called *swamping argument* (see Elga 2007, sect.9; and T. Kelly 2011, sect.3): This argument stresses the fact that according to (EWV) first order evidence can be completely swamped by higher order evidence. However, at least at first sight this seems to be implausible: It seems that one's own arguments should not count for nothing in case of an increasing number of peers who disagree. Indeed, as we have mentioned already in the description of figure 10.1, if the number of peers increases, the influence of first order evidence decreases. However, it is important to note that it is not any number of contestants that has to increase, but the number of *peer contestants*. The general method we have introduced in the preceding section, the meta-inductivist Pr_m, is a social strategy. And as a social strategy it has to take her peers seriously. As one can see according to equations (10.2) to (10.4), this method is even a completely social strategy inasmuch as it bases her inferences not at all on any individual argument or first order evidence. It just operates on higher order evidence. And as the optimality result regarding Pr_m as well as the suboptimality results regarding (RSV) and (TEV) show, taking into account higher order evidence only is not just a sufficient means for achieving long run optimality, but also a necessary one. In this sense our framework completely supports Adam Elga's view, who claims: "If one really has 99 associates who one counts as peers who have independently assessed a given question, then one's own assessment should be swamped" (Elga 2007, p.494). Regarding optimality considerations, our framework shows why and how higher order evidence overwrites first order evidence: Inferences based on first order evidence are prone to being deceived, whereas inferences based on higher order evidence drastically decrease possibilities of being deceived relative to one's fellows.

This holds for cases of disagreement amongst peers. Regarding disagreement amongst non-peers, there is also a conclusion one can draw from our framework: The long run optimality result for Pr_m is based on the proof of an upper bound for differences in the reliability between the individual inferences and that of the social method. As is shown in theorem 3.40, for Pr_m the following bound holds with respect to any inference method Pr_i of $Pr_1, \ldots, Pr_n$:

$$s_i^t - s_m^t \leq \sqrt{n/t}$$

In the limit this term goes to 0 which means that s_m approaches also the reliability of the best agents in the setting. However, for the short run, i.e. t not arbitrarily high, n, i.e. the number of agents in the setting, has a relevant influence on the performance. The more non-peers there are in the setting, the more the social strategy is prone to errors. So, in such cases individual arguments and first order evidence might easily overwrite higher order evidence, i.e. higher order evidence should not swamp first order evidence. Notice, however, that this concerns cases of disagreement amongst non-peers. And for these cases (EWV) leaves the choice of weights for incorporating first and higher order evidence completely open.

A further common objection put forward against (EWV) is that of *spinelessness* of equal weighting (see Elga 2007, sect.9): The idea of this argument is that according to (EWV) too often a response to disagreement will be something close to suspension of judgement on virtually all controversial issues. Elga argues that this is probably not that often the case as one might expect, because cases of *real* epistemic peer disagreement are not that widespread (see Elga 2007, p.495). Regardless of this, our model shows that even if epistemic peer disagreement is a widespread phenomenon, from an optimality point of view something close to suspension of judgement might turn out to be the optimal strategy. However, this brings us also to an important assumption made in our model: In the reliability measure of equation (10.1) we have assumed a quadratic difference measure between the truth value of the proposition in question and the inference prediction by the agent. The optimality result for Pr_m and (EWV) can be also achieved by any other *convex* distance measure. However, no such general result is proven for non-convex distance measures. If one wants to make the optimality of an inference method dependent on its being decisive regarding the extreme values 0 and 1 (and penalises suspension and close to suspension cases), then one might want to use a non-convex distance measure. Such a measure might assign linear weights to distances between 0 and 0.25 and maximal weight to distances above 0.25. Figure 10.3 provides an example of such a measure. It is important to note that for such non-convex distance measures our argument does not privilege (EWV) compared to (RSV) and (TEV). This clearly is a restriction of our model.

One nice feature of our model is that it allows for a natural explanation of the traditional arguments for (EWV) by help of symmetry considerations: As we have already mentioned in the introduction, sometimes a principle we called the *principle of epistemic indifference* is stressed for equal weighting. Our framework does not presuppose such a principle. Rather, it follows from the optimality considerations of our framework: The idea is that once we aim at optimality with respect to predictions and inferences, one can prove that epistemic indifference is the right principle to apply: If the epistemic performance (reliability) of n agents is indistinguishable, i.e. they are peers, then their epistemic attitudes need to be weighted equally in order to achieve an optimal inference or prediction.

We can sum up the cornerstones of our investigation in this chapter as

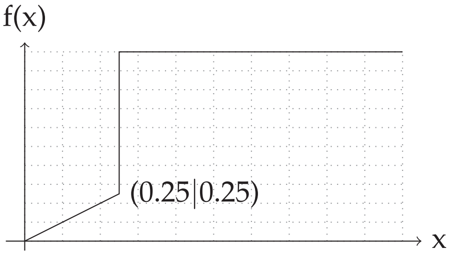

Figure 10.3: Example of a non-convex distance measure: The reliability of an inference (see equation (10.1)) is measured via $1 - f(|val(p) - Pr(p))|$, where distances within $[0, 0.25]$ are measured linearly, whereas distances within $(0.25, 1]$ are measured maximally. By this epistemic attitudes that are far away from the truth or close to suspension of judgement are penalised maximally, i.e. one's reliability decreases maximally. The employed optimality result does not hold for such a non-convex distance measure.

follows: In the debate of epistemic peer disagreement the central question concerns the problem of how to incorporate higher order evidence about epistemic peers disagreeing with one's epistemic attitudes. In this chapter we have presented a model for such disagreement that frames this problem as a problem of updating one's credences given such higher order evidence. We have identified the conditions for epistemic peers in (i) updating on one and the same set of evidence, and (ii) in having the same inferential skills regarding this evidence, measured by a reliability track record. We were able to define in this model the main three traditional approaches: the remain steadfast view (RSV) which suggests to ignore such evidence in updating by assigning it zero weight. The equal weight view (EWV) which suggests to update on such evidence by equally weighting it. In case the number of peers increases, higher order evidence has much more impact than first order evidence and can even swamp it. And finally, the total evidence view (TEV) which suggests to consider higher order evidence simply as just another kind of evidence. Since updating on higher order evidence is a social epistemological action, we have put forward an optimality constraint for such actions by aiming at optimality: One's update on higher order evidence should be such that (in the long run) one's inferences are optimal in the sense that they are the most reliable ones compared to the other inference methods of the setting. By employing the framework of *meta-induction* we were able to show that (EWV) satisfies this optimality constraint for cases of epistemic peer disagreement, whereas (RSV) and (TEV) fail to do so. This adds a new argument to the debate on epistemic peer disagreement which seems to be decisive with respect to (EWV).

Chapter 11

Judgement Aggregation

There is a plurality of formal constraints for aggregating probabilities of a group of individuals. Different constraints characterise different families of aggregation rules. This chapter focuses on the families of linear and geometric opinion pooling and shows that applying the main optimality result of the theory of meta-induction provides a general rationale for choosing the weights in a success-based way by scoring rules.

Probability aggregation is the theory of how to adequately aggregate several probability distributions to a single one. For more than two decades now disciplines concerned with probabilistic reasoning and its rationale are undergoing a *social turn*, at least so it seems. This makes the problem of probability aggregation a highly relevant topic. So, e.g., in philosophy of science recent research centres not only around the relation between evidence, theory, and explanation, which is quite often spelled out in probabilistic terms, but also on the relations between scientific groups having different evidence, holding different theories, and providing alternative explanations (see Douven and Riegler 2010; Hartmann, Martini, and Sprenger 2009; Zollman 2007). Similarly in epistemology, where—as we have seen previously—core topics of social epistemology like testimony, peer disagreement, and judgement aggregation are very often framed in a probabilistic setting (see, e.g., for testimony Goldman 1999; for peer disagreement Elga 2007; and for probabilistic judgement aggregation Dietrich and List 2016). It is clear that also here the question of how to aggregate one's own and one's testifier's or peer's probabilistic opinions shows up.

In a similar line as it is argued in social choice theory, also in the theory of probability aggregation general constraints for such aggregation methods are put forward; subsequently it is investigated which aggregation methods satisfy these constraints. Often the constraints put forward are not compatible with each other. This led to the famous impossibility results of social choice theory (see Arrow 1963) and the theory of judgement aggregation (see List and Pettit 2002). However, as it turned out, one can cluster these constraints in such a way that relevant subclasses are jointly satisfiable and characterise different families of aggregation methods. As we will see in the next section, broadly accepted constraints lead in particular to two common aggregation rules, namely *linear weighting* and *geometric weighting*. So, if one can figure out which constraints for probability aggregation are relevant for which domain of application, one seems to be able to give a partial solution to the problem of probability aggregation. However, even if one subscribes to such a purpose-dependent strategy (see, e.g., List and Pettit 2011), the constraints put forward at most determine a *family* of aggregation methods, but no *exact* aggregation method. In particular, the choice of the weights—which is from the viewpoint of practical applications the most important factor—is undetermined by these constraints.

In this chapter we are going to argue for a new approach to determine such weights. We will do so by employing the main optimality result of meta-induction and show that a success-based determination of weights allows for proving long run optimality of probabilistic predictions. This allows, on the one hand, for a more specific determination of the weights used for aggregating probabilities, and on the other hand it also provides an epistemic rationale for doing so.

The structure of this chapter is as follows: In section 11.1 we summarise the characterisation results of the theory of probability aggregation which lead to two families of aggregation functions, namely the linear and the geometric weighting rules. Since the exact weights are not determined by these results we briefly discuss solutions to determining weights and their problems in section 11.2. There we also outline our solution and present an instance of the framework of prediction games which we want to employ for a justified choice of weights for probability aggregation. This prepares the ground for section 11.3, where we apply this framework to a probabilistic setting: We show how the meta-inductive optimality results can be transformed to the probabilistic forecasters and provide a general epistemic rationale for success based linear probability aggregation. Finally, in section 11.4 we show how a more restricted result can be also employed in geometric probability aggregation.

11.1 Probability Aggregation and Characterisation Results

Many investigations on probability aggregation were triggered by Leonard J. Savage's seminal work on the *Foundations of Statistics*, where he introduced a "model of group decision". He described the problem at hand as follows:

> "Consider a group of people [. . .] supposed to have the same utility function, at least for the consequences to be considered in the present context, but their personal probabilities are not necessarily the same. The group of people is placed in a situation in which it must, acting in concert, choose an act [. . .] from a finite set of available acts [. . .]. The situation just described will be called a *group decision problem*." (see Savage 1972, chpt.10.2)

A paradigmatic example mentioned by Savage is the decision making by a jury in a court of law: Such a jury is supposed to have common value judgements for these are incorporated in the law as stated in the instructions of the court. But, on the other hand, it is clear or desirable of a jury that its members may have different opinions. Still, as the jury has to come to a conclusion as a jury, it needs to end up with a group opinion and often a unanimous decision. The scheme of the problem is as follows:

$$Pr_{\{1,\dots,n\}} = f(Pr_1, \dots, Pr_n)$$

Here $Pr_1, \dots, Pr_n$ are the graded opinions, probabilities, credences or graded predictions of the members of a group, f is an aggregation function, and $Pr_{\{1,\dots,n\}}$ is the respective graded group opinion, probability, credence or graded prediction. In what follows we assume that the set of propositions under investigation is a finite set of k primitive and mutually independent propositions (also called *propositional variables*); this gives us $\mu = 2^k$ constituents of the form $c_i = (\pm)p_1\& \dots \&(\pm)p_k$; we refer to these constituents simply as "possible worlds" (as they represent possible worlds—we avoid speaking of "atomic propositions" or "atoms" because usage of this notion is ambiguous: linguistically one meaning is that of "primitive" propositions and algebraically one meaning is the notion of a possible world; obviously these two notions are entirely different). In later sections the p_is will represent primitive propositional functions with a time variable t.

How to constrain the transmission from the individual to the group is considered to be the *group decision problem*. Savage discussed and criticised two constraints: First, that such a transmission has to be in accordance with the minimax rule, i.e. that the largest *expected* loss faced by any member of the group has to be as small as possible. Note that here 'expected' refers to the individual expectations (before choosing an aggregated probability distribution), not to that of the group (see Savage 1972, p.174). And second, that some kind of *Pareto principle* has to be satisfied, the so-called *principle of admissibility*, i.e. that in case the minimax rule allows for different aggregated probability functions, no outcome should be chosen that is outperformed in terms of the expected utilities of all individuals by some other outcome that still satisfies *minimax*.

More generally, a plurality of constraints for the group decision problem has been widely discussed. Such investigations are often performed in the line of the so-called *axiomatic method*, where one does not choose a particular aggregation function directly, as, e.g., a linear aggregation function, and then proves some properties of it, but instead one formulates general constraints on a *good* aggregation function in the form of axioms, and then asks which aggregation functions satisfy them, if any at all (see Dietrich and List 2016, sect.3). A vast amount of literature evolved in this area which was collected and commentated by Genest and Zidek (1986).

The axiomatic method has a long tradition also in social choice theory. Seminal is Kenneth Joseph Arrow's *Social Choice and Individual Values* in which he was able to prove an infamous *impossibility result* of opinion aggregation already as early as 1951. An error in the statement of these results is corrected in the second edition of 1963 which is the main reference now. However, the predominant framework of this domain is not that of probability aggregation, but that of judgement and preference aggregation, which is typically within a nominal or ordinal scale. Contrary to this, probability aggregation operates on a cardinal scale. As is often the case, impossibilities on nominal and ordinal scales disappear on a cardinal scale. Nevertheless, Arrow's investigations of the former realm were also taken as a model for investigations of impossibility and characterisation results about the latter one (e.g. also Savage refers to Arrow).

The motivation put forward for the axioms used in this approach is manifold. To mention the most prominent reasons, we provide here a short list categorising constraints according to their cross disciplinary justification (see Gaertner 2009; and Gaertner 2016):

- Arguments from *social ethics* in favour of: minimax rule, non-dictatorship, anonymity (permuting the credences among individuals should not influence the outcome), universal domain (no credence function should be excluded on a priori grounds), the possibility of Paretian liberals (it should be possible that there are individuals who are decisive for the group credence regarding a "private sphere" of propositions), the possibility of alienable rights (it should be possible to provide hierarchies for individual credences to allow an individual to waive her credence' impact), avoiding the no-show paradox (not considering individual credences in favour of a proposition should not allow for increasing the group credence in that proposition), strategy proofness (no individual should be able to increase her gains from the group opinion in a group decision problem by misrepresenting her true opinions; we will discuss this constraint later on in the context of *proper* scoring rules)

- Arguments from *optimisation* in favour of: Pareto principles (e.g. the constraint to preserve unanimous opinions), the maximax rule (the largest gains faced by any individual should be as big as possible), Condorcet consistency (if there is a Condorcet winner—i.e. a candidate which wins pairwise in comparison with all the other candidates—, then it is also elected; regarding propositions: If there is a proposition which gets unanimously highest credence, then also the group credence should be highest), avoiding Condorcet losers (analogous to Condorcet winners)

- *Coherence* arguments in favour of: non-cyclicity (in comparative versions of the group decision problem no cycles should show up), partition consistency or reinforcement (splitting up the jury, aggregating partially, and combining the partial aggregation result should not change the outcome), monotonicity or positive responsiveness (increasing an individual's credence in a proposition should, everything else being equal, not decrease the group credence in that proposition)
- Irrelevance of independent alternatives (the group opinion on a proposition should be dependent only on individual opinions regarding *that* proposition)
- Etc.

Many impossibility results of subsets of the constraints indicated above have been proven in the past. As we mentioned already, seminal is, e.g., (Arrow 1963), where it is shown that four very basic constraints cannot be simultaneously satisfied in the comparative realm: If there are at least three alternatives and individuals, then the axioms (for details see below) of *universal domain*, the *weak Pareto principle*, and *independence of irrelevant alternatives* imply *dictatorship* and by this *non-anonymity* (see Arrow 1963, p.97). (For propositions in terms of *ranking theory* the weak Pareto principle might be expressed as follows: $\forall i \in \{1,\ldots,n\} : \kappa_i(p) \leq \kappa_i(q) \Rightarrow \kappa_{\{1,\ldots,n\}}(p) \leq \kappa_{\{1,\ldots,n\}}(q)$.) List and Pettit (2002) prove a similar result for the qualitative realm of opinions, namely belief and disbelief. In this realm the *weak Pareto principle* is not necessary for an impossibility theorem, instead further formal constraints are needed (see below). Also here problems of the qualitative and comparative realm disappear on the quantitative one. What is more, the three axioms that together with other plausible constraints (see below) lead to an impossibility result within the qualitative realm even characterise a plausible family of transformations or aggregation rules. As is discussed and shown in (Lehrer and Wagner 1981, chpt.6; and Genest and Zidek 1986, sect.3), the mentioned three conditions:

(U) *Universal domain*: The Pr_i satisfy the laws of probability theory ($\forall i \in \{1,\ldots,n\}$)

(A) *Anonymity/Permutation*: $Pr_{\{1,\ldots,n\}} = f(Pr_1,\ldots,Pr_i,Pr_{i+1},\ldots,Pr_n) = f(Pr_1,\ldots,Pr_{i+1},Pr_i,\ldots,Pr_n)$

(I) *Irrelevance of Alternatives*: There is not only a function f such that $Pr_{\{1,\ldots,n\}} = f(Pr_1, \ldots, Pr_n)$, but there is also a function f^* such that $Pr_{\{1,\ldots,n\}}(p) = f^*(Pr_1(p), \ldots, Pr_n(p))$ (for all propositions p)

characterise the family of linear opinion aggregation rules which have the form of a weighted arithmetic mean:

$$Pr_{\{1,\ldots,n\}} = \sum_{i=1}^{n} w_i \cdot Pr_i \qquad \text{(where } w_i \geq 0 \text{ and } w_1 + \cdots + w_n = 1) \tag{AM}$$

So, what resulted in problems and questions of choosing among fundamental formal constraints for modelling a group's opinion before, turns out to determine an important family of functions now, namely linear opinion aggregation rules. Since many theorists consider (U), (A), and (I) to be plausible constraints for probability aggregation, this family has been also proposed as a general framework for probability aggregation (see Lehrer and Wagner 1981).

However, this characterisation also has some problems. One important drawback is that (U), (A), and (I) are jointly incompatible with other further plausible constraints for aggregating probabilities. Well known is, e.g., their incompatibility with the axiom of *independence preservation* (see Lehrer and Wagner 1983): This axiom demands that if all members of a group consider two propositions to be probabilistically independent: $Pr_i(p_1|p_2) = Pr_i(p_1)$ $(\forall i \in \{1, \ldots, n\})$, then also the aggregation should be this way: $Pr_{\{1,\ldots,n\}}(p_1|p_2) = Pr_{\{1,\ldots,n\}}(p_1)$. Connected with this is the problem that the constraint of aggregating *externally Bayesian* (see Genest and Zidek 1986, p.119), also called *commutativity with learning* (see Brössel and Eder 2014), is not compatible with these conditions: Aggregating individual credences and then performing a Bayesian update by new evidence might be different from all individual's first performing a Bayesian update of their credences and then aggregating the updated credences (see Mongin 2001, p.320). The commutative update rule resulting from linear weighting is called "imaging" and differs in important respects from Bayesian updating (see Leitgeb 2016). However, there is another family of aggregation functions that allows one to satisfy the commutativity constraint while still upholding Bayesian orthodoxy: Genest (1984, p.1101) and Genest, McConway, and Schervish (1986, p.499) show that weak unanimity preservation (see Russell, Hawthorne, and Buchak 2015, p.1295, fn.8) and external Bayesianity and some further technical assumptions characterise the family of the logarithmic or *geometric* graded opinion aggregation rules. For lack of space we will not discuss the technical assumptions here. However, the constraints of weak unanimity preservation and external Bayesianity can be characterised easily in detail:

(P) *Weak Unanimity Preservation*: If $Pr_1 = \cdots = Pr_n$, then $Pr_{\{1,\ldots,n\}} = Pr_1 = \cdots = Pr_n$.

(B) *External Bayesianity*: There is a function f^* such that for all propositions p, q: $Pr_{\{1,\ldots,n\}}(p|q) = \frac{Pr_{\{1,\ldots,n\}}(p\&q)}{Pr_{\{1,\ldots,n\}}(q)} = \frac{f^*(Pr_1(p\&q),\ldots,Pr_n(p\&q))}{f^*(Pr_1(q),\ldots,Pr_n(q))}$

These constraints characterise the normalised weighted geometric mean as defined below. Note that Franz Dietrich and Christian List think that the additional technical assumptions needed for proving a characterisation result are in need of further justification for which reason they advance the view that "we still lack a fully compelling axiomatic characterization of geometric pooling" (Dietrich and List 2016, sect.6).

The normalised weighted geometric mean of a family of probability functions is defined as follows:

$$Pr_{\{1,\ldots,n\}}(c_l) = \frac{\prod\limits_{i=1}^{n} Pr_i(c_l)^{w_i}}{\sum\limits_{j=1}^{\mu}\prod\limits_{i=1}^{n} Pr_i(c_j)^{w_i}} \qquad \text{(GM)}$$

(where Pr_i is regular, $c_l \in \{c_1, \ldots, c_\mu\}$,
$w_i \geq 0$ and $w_1 + \cdots + w_n = 1$)

This family of aggregation rules is technically quite demanding. The denominator in the equation above guarantees $normalisation(\top) = Pr_{\{1,\ldots,n\}}(c_1 \vee \cdots \vee c_\mu) = Pr_{\{1,\ldots,n\}}(c_1) + \cdots + Pr_{\{1,\ldots,n\}}(c_\mu)$. Since the set of worlds is supposed to be finite, the equation above determines $Pr_{\{1,\ldots,n\}}$ for arbitrary propositions, i.e. disjunctions of possible worlds, via $Pr(c_i \vee c_j) =_{def} Pr(c_i) + Pr(c_j)$.

We should mention, however, that geometric aggregation suffers from an oddity compared to arithmetic aggregation. If the arithmetic aggregation rule is defined for possible worlds, then it is preserved for arbitrary propositions, i.e., it applies also to disjunctions of possible worlds (as can be easily proved). This is not so for geometric aggregation, which applies only to possible worlds, but not to disjunctions of them; rather, the geometrically aggregated probability of a disjunction of worlds is defined, as explained above, as the sum of the aggregated probabilities of these worlds (see Dietrich and List 2016, sect.6). For example, if a given proposition is the disjunction of two worlds (constituents) $c_1\dot{\vee}c_2$ and the probabilities of the worlds of the two experts are $Pr_1(c_1), Pr_2(c_1)$ and $Pr_1(c_2), Pr_2(c_2)$ respectively, then the geometrically aggregated probability of the two worlds are $Pr_1(c_1)^{w_1} \cdot Pr_2(c_1)^{w_2}$ and $Pr_1(c_2)^{w_1} \cdot Pr_2(c_2)^{w_2}$ respectively and the aggregated probabilities of the disjunctive event $c_1\dot{\vee}c_2$ is by definition $Pr_1(c_1)^{w_1} \cdot Pr_2(c_1)^{w_2} + Pr_1(c_2)^{w_1} \cdot Pr_2(c_2)^{w_2}$, which in general is different from the result of applying geometric aggregation to the disjunctive proposition which is $(Pr_1(c_1) + Pr_1(c_2))^{w_1} \cdot (Pr_2(c_1) + Pr_2(c_2))^{w_2}$.

That the Pr_is are *regular* means that the individual probability of every possible world is greater than 0, i.e.: $Pr_i(c_l) > 0$ for all $i \in \{1, \ldots, n\}$ and every $c_l \in \{c_1, \ldots, c_\mu\}$. This assumption is necessary, because otherwise it is possible that for each possible world there is an individual whose probability or credence of that world is 0; in this case the generalised geometric mean would be zero too, which would be probabilistically inconsistent.

More details of the family of geometric aggregation rules are discussed, e.g., in (Dietrich and List 2016, sect.6). Regardless of the exact characterisation of arithmetic and geometric aggregation rules and the assessments of their advantages and disadvantages, these two families are amongst the most common pooling methods. And, although there is no general aggregation method that allows one to satisfy the constraints for aggregating probabilities as put forward here simultaneously, these two families allow one to satisfy reasonable subsets of these constraints. If one follows the line of argumentation of List and Pettit (2011) and makes the choice of the exact aggregation rule dependent on the context and purposes in question, then (AM) and (GM) may seem to be good candidates for solving the group decision problem. Hence, it should not make one wonder too much that these two families are also the two most prominent types of probability aggregation rules studied in the literature.

However, there is a problem underlying both (AM) and (GM): It is true that the characterisation results make clear which axioms determine the choice of which family of aggregation rules. Nevertheless, each family still allows for a wide range of different aggregations. And as one can easily see when looking at the equations, this variance is due to the underdetermination of the weights by the aggregation constraints. So, in order to provide an adequate answer to the group decision problem, one also has to address the problem of choosing the right weights.

11.2 The Problem of the Underdetermination of the Weights

As we have indicated above, the constraints (U), (A), and (I) determine the *family* of linear aggregation rules, (P), and (B) (and some technical assumptions not described here further) determine the *family* of geometric aggregation rules, but no set of the constraints allows one to determine a *specific* aggregation rule. Regarding the weights used for aggregation these constraints remain undetermined. Now, it is sometimes suggested in the literature that there is no general objective account of justifying a specific choice of the weights: “The determination of the weights is a subjective matter, and numerous interpretations can be given to the weights” (Clemen and Winkler 2007, p.157). Also Genest (1984, p.1104) mentions this problem when stating his characterisation result of (GM): "The problem of choosing the weights w_i [...] remains and is not addressed here. This difficulty is common to most axiomatic approaches".

Genest and McConway (1990) provide an overview of approaches to determine weights and briefly discuss their problems. We are going to mention just the most prominent approaches here.

According to the interpretation of *veridical probabilities* (see Bunn 1981, p.213), weights are considered to represent the probability of an individual probabilistic forecast to be right: "w_i represents the probability that Pr_i is the 'true' distribution" (see Genest and McConway 1990, p.56, notation adjusted) and "w_i would represent the probability of predictor i being the 'true' descriptive model of the underlying stochastic process" (see Bunn 1981, p.213, notation adjusted). So, according to this approach the weights w_i represent the "decision maker's" credence in Pr_i making an accurate prediction: $Pr_{\{1,\ldots,n\}}(Pr_i = ch)$, where ch is the true chance distribution (see Bunn 1981, p.213). However, this approach faces the main problem that it is not clear how one can determine the relative veracity of competing opinions when one is entirely ignorant about the true distribution in the world. Moreover, at any stage of evidence this account faces the problem of induction, i.e. of estimating the distribution over unobserved individuals from the observed individuals; and different *priors* give entirely different anwers to this problem. In conclusion, the account fails to tell us what should be considered as adequate priors of $Pr_{\{1,\ldots,n\}}$ in estimating that Pr_i is an accurate distribution. For this reason, so it seems, this approach fails to set foot on solid ground.

It was also suggested to consider only such weights that *minimise variance* between the group opinion and the individual opinions: If the individual's Pr_is are unbiased, then "weights [should] be chosen so as to minimise the variance of [$Pr_{\{1,\ldots,n\}}$], the composite forecast" (see Genest and McConway 1990, p.60). That the Pr_is are unbiased means that for all possible worlds c_l the expected difference between $Pr_i(c_l)$ and the objective chance of c_l equals 0 (where c_l is a "state description" or a possible value of a repeatable event). The rationale behind this approach is that a certain linear combination of unbiased probability distributions decreases the error variance. For example, if there are unbiased predictors Pr_i that have independent forecast error variances σ_i^2, then combining the probability distributions by the following linear weights:

$$Pr_{\{1,\ldots,n\}}(c_l) = \sum_{i=1}^{n} \frac{\sum\limits_{i \neq j \in \{1,\ldots,n\}} \sigma_j^2}{2 \cdot \sum\limits_{j=1}^{n} \sigma_j^2} \cdot Pr_i(c_l)$$

leads to $Pr_{\{1,\ldots,n\}}$ having a smaller or equal forecast error variance $\sigma^2_{\{1,\ldots,n\}}$

than the Pr_is, i.e. $\sigma^2_{\{1,\ldots,n\}} \leq \sigma^2_i \; \forall i \in \{1, \ldots, n\}$ (see Bunn 1981, p.213). A problem of this interpretation is that it not clear how to technically expand it also to cases where individuals' graded predictions are biased.

In a further approach the weights are interpreted as *outranking probabilities*: "w_i should be interpreted as the probability that the next prediction made using opinion Pr_i will outperform predictions made from all other individual opinions in the group" (see Genest and McConway 1990, p.57, notation adjusted). An advantage of this interpretation is that such weights are operationally easier to grasp. "However, the main problem with this approach is that if the experts know in advance how their weights will be derived, they may experience them as scores and choose to report dishonest opinions in order to maximise their influence on the opinion pool". This was the reason for introducing another interpretation of the weights, namely weights being interpreted as *scores*: In order to avoid the problem of manipulation, *proper* scoring rules for weights were put forward, i.e. scoring rules which guarantee "that the distribution reported by each expert maximises his expected utility if he is honest and coherent". However, also here a problem seems to show up: There is a plurality of proper scoring rules (quadratic, logarithmic, spherical etc.) and empirical investigations suggest that "weights [resulting from scores are not] quite satisfactory because they seemed sensitive to the choice of scoring rule" (see Genest and McConway 1990, pp.56ff).

This is the point where we think that meta-induction should enter the picture, because it allows for determining weights generally in a success-based way. Then optimality results of meta-induction can be cashed out for providing a general rationale for such a determination. The main line of our argumentation is that at least for linear pooling the epistemological rationale provided by the optimality result of meta-induction is general enough to capture all relevant scoring rules. So, in order to accommodate this rationale, no specific choice of a scoring rule is necessary. Rather, many of them can be justified generally and the exact choice of a scoring rule might be plausibly made dependent also on the context and purpose in question.

In the remainder of this section we list the optimality results of the theory of meta-induction which we are going to employ for a meta-inductively justified choice of weights for probability aggregation in the subsequent sections.

Recall the main ingredients of our setting, namely prediction games (see chapter 2):

- $Y_1, Y_2, \ldots$ is an infinite series of events (variables) whose outcomes (values) $y_1, y_2, \ldots$ are elements of the normalised interval $[0, 1]$
- $f_{1,t}, \ldots, f_{n,t} \in \mathcal{F}$ are the predictions of Y_t (also elements of $[0, 1]$) of all n accessible prediction methods, the so-called *candidate* methods,

which are typically but not necessarily object-level methods. Since we are interested in a probabilistic interpretation we will use for the elements of $\mathcal{F}$ throughout this chapter:
$Pr_{1,t}, \ldots, Pr_{n,t}$

- $Pr_{mi,t}$ is the prediction of Y_t of the meta-inductive method (this is the learning algorithm defined on $\mathcal{F}$ of chapter 3)

As we have seen in section 5.4, a meta-inductive method "cooks up" a prediction from the present predictions and past success rates of the candidate methods. In order to keep track of the success rate of a method i one measures the score of i's prediction about event Y via measuring 1 minus the negative loss ℓ of its predictions for each round and then summing all of its scores up to event Y_t and dividing by t:

$$succ_{i,t} = \frac{\sum\limits_{u=1}^{t} 1 - \ell(Pr_{i,u}, y_u)}{t}$$

The measure $succ_{i,t}$ represents the average per-round success rate of candidate method i up to prediction round t. Recall, the only assumption we make about the loss measure ℓ is that it is within $[0, 1]$, and that it is *convex* in its first argument, i.e. that the loss of a weighted average of two predictions is lower or equal to the weighted average of the losses of these two predictions. Or formally: $\ell(w \cdot x + (1 - w) \cdot y, z) \leq w \cdot \ell(x, z) + (1 - w) \cdot \ell(y, z)$ holds for all x, y and $w \in [0, 1]$.

Now, based on this measure for the success rate up to round t we define the weights via the relative success or *attractivities*:

$$w_{i,t} = \frac{max(0, succ_{i,t} - succ_{mi,t})}{\sum\limits_{j=1}^{n} max(0, succ_{j,t} - succ_{mi,t})}$$

Note that the weights are positive and sum up to 1; prediction methods that are performing worse than Pr_{mi} get weight 0. If Pr_{mi} outperforms all candidate methods, then $succ_{mi,t} \geq succ_{i,t}$ for all $i \in \{1, \ldots, n\}$, and we stipulate $w_{i,t} = 1/n$.

Based on these weights, we define the weighted-average meta-inductive method:

$$Pr_{mi,t+1} = \sum_{i=1}^{n} w_{i,t} \cdot Pr_{i,t+1} \tag{AMI}$$

Recall, we can also define such weights in the exponent, simply by defining

$$ew_{i,t} = \frac{e^{-\eta \cdot \sum_{u=1}^{t} \ell(Pr_{i,u}, y_u)}}{\sum\limits_{j=1}^{n} e^{-\eta \cdot \sum_{u=1}^{t} \ell(Pr_{j,u}, y_u)}}$$

The exponential success-dependent meta-inductive predictor is defined similarly to the linear success-dependent meta-inductivist (AMI) by the method of weighted arithmetic average; thus:

$$Pr_{emi,t+1} = \sum_{i=1}^{n} ew_{i,t} \cdot Pr_{i,t+1} \tag{EAMI}$$

Both methods allow for relative learnability: The bounds of Pr_{mi} and Pr_{emi}'s relative worst-case regret, i.e., the loss of their success rates compared to the success rate of the actual best candidate method are (see the proofs in section 3.4):

- Given (AMI), $succ_{i,t} - succ_{mi,t} \leq \sqrt{n/t} \quad (\forall i \in \{1, \ldots, n\})$.
- Given (EAMI), $succ_{i,t} - succ_{emi,t} \leq \frac{2}{\sqrt{2}-1} \cdot \sqrt{ln(n)/t} \quad (\forall i \in \{1, \ldots, n\})$.

As proven here, if $n \geq 110$ the exponential success-dependent meta-level method (EAMI) has a better guaranteed lower bound (there are better proofs of better bounds for equation (EAMI) which show an advantage already with $n > 6$ in the literature (see Schurz 2019, sect.6.6.2)). It should be noted also that (EAMI) is the best known long run access optimal meta-inductive method inasmuch as it best approximates the minimal lower bound that is achievable in principle, namely $\sqrt{\ln(n)/2t}$ (see Cesa-Bianchi and Lugosi 2006, p.62, thrm.3.7). However, what is most important in our context is that the relative regret of the two meta-inductive methods converges quickly to zero when t grows large. An important consequence of this fact is the following result on the so-called *long run acccess-optimality* of meta-induction:

- Given ℓ is convex (where ℓ is used for determining s), then both meta-inductive prediction methods (AMI) and (EAMI) are optimal in the long run:

$$\begin{aligned} \lim_{t\to\infty} max(succ_{1,t}, \ldots, succ_{n,t}) - succ_{mi,t} &\leq 0 \\ \lim_{t\to\infty} max(succ_{1,t}, \ldots, succ_{n,t}) - succ_{emi,t} &\leq 0 \end{aligned}$$

In the next sections we are going to utilise these results in order to determine the weights of linear and geometric probability aggregation and provide an epistemic rationale for such a determination.

11.3 Meta-Inductive Linear Probability Aggregation

In probabilistic prediction games each forecaster or candidate method identifies the predicted real value with its credence of the predicted event conditional on her information about the past. First, let us ask: When is it reasonable to equate one's real-valued prediction with one's probability of the predicted event? According to a well-known result, this identification is not optimal if the loss function is natural or linear, even if one's probability is close to the true statistical probability. Rather, under this assumption the optimal prediction rule is the so-called *maximum rule* which predicts that event value v whose conjectured probability (i.e., so-far observed frequency $freq_t$) is maximal (see Rumelhart and Greeno 1971; Reichenbach 1938, pp.310f). Recall, in section 9.4 we discussed this rule already (under the label *take-the-most-frequent strategy*). For binary events the maximum rule predicts 1 as long as $freq_t(1) \geq .5$ and 0 otherwise.

Now, although with a linear loss function the maximum rule is better off in predicting the probabilities of discrete events, we are still often interested in one's full probabilistic estimations (arguments for an agreement between epistemic/subjective and objective probabilities are discussed, e.g., in Schurz 2019, sect.7.1). In order to enforce a forecaster to reveal her *real* credences, non-linear scoring rules have been devised. With such scoring rules the expected success of real-valued predictions in independent and identically distributed sequences (*i.i.d.*, see definition 2.18) is maximal exactly if the forecaster predicts according to her credences. The relevant property for such loss functions is the following one (see Schurz 2019, statement 7.1):

(PS) *Proper Scoring*: The loss function ℓ for a *proper scoring rule* of a binary event Y with event outcome y has to satisfy the following constraint: The expected loss of prediction r under probability Pr, $Exp_{Pr}(r) =_{def} Pr(y=1) \cdot \ell(r,1) + Pr(y=0) \cdot \ell(r,0)$, is minimal iff $r = Pr(y=1)$.

Now, given such a loss function, clearly there is an incentive for a forecaster to predict according to her credences: if she does so, then she maximises her expected success. Linear losses fail to satisfy requirement (PS); however, certain non-linear but convex loss functions are in agreement with it. So, e.g., Brier (1950) showed that the quadratic loss function (with $\ell(r,y) = (r-y)^2$) is a proper scoring rule (for details on this see Schurz 2019, sect.7.1). The (expected) quadratic loss function is a standard measure in statistical investigations of, e.g., judgement aggregation. However, it was criticised due to penalising large deviations to a much higher degree than small ones—there is also empirical evidence that decision-makers are decreasing sensitivity with increasingly larger deviations from the true value (see Hartmann and Sprenger 2010, p.347, there also conditions for more "realistic loss functions" are discussed). As we will see below, meta-inductive probability aggregation works (is optimal) when using a quadratic loss function, but it works also with many other (convex) loss functions.

In the following part of this section as well as in section 11.4 we discuss implementations of meta-induction into the framework of probability aggregation. We will start with an implementation which allows for proving general optimality for linear pooling. By this, e.g., the quadratic loss function proposed by Brier (1950) is proven to be optimal. Then we will go on with proving a much more restrictive optimality result for the much more complicated case of geometric pooling. Although scoring functions satisfying constraint (PS) seem to be the most adequate ones for probabilistic forecasts, the following considerations will hold for all convex scoring functions.

In order to cash out the long run access optimality result of meta-induction presented above for probability aggregation we have to change our framework: It contains:

- Again, a series of events represented by random variables $Y_1, Y_2, \ldots,$ but now the events do not have outcomes within $[0, 1]$, but within a space of discrete (non-numerical), mutually disjoint and exhaustive values v_m, $Val = \{v_1, \ldots, v_k\}$. In order to indicate which value a random variable took on at a specific round t, we assume a valuation function val_t to be given by:

$$val_t(v_m) = \begin{cases} 1, & \text{if the value of } Y_t \text{ is } v_m \\ 0, & \text{otherwise} \end{cases}$$

- Predictions are the credences of n candidate methods for each event variable Y_t in the series, represented by probability distributions $Pr_1, \ldots, Pr_n$:

$$\forall\, t, i \in \{1, \ldots, n\} \sum_{m=1}^{k} Pr_{i,t}(v_m) = 1 \text{ (and } Pr_{i,t}(v_m) \geq 0)$$

  So, for each event, at each round, the candidate methods provide a full probability distribution about the outcome of the event in question.

- The success-based meta-inductive methods Pr_{ami}, Pr_{gmi} are also represented by a probability distribution and defined as an arithmetically/geometrically weighted average of the $Pr_1, \ldots, Pr_n$; details are presented below.

The attempt to expand the meta-inductive framework of prediction games to the probabilistic setting faces the problem that the predictions are real numbers, i.e. probabilities, but the event's values are not numbers but non-numeric mutually exclusive and exhaustive values $v_1, \ldots, v_k$. There are different possibilities to apply the meta-inductive framework of prediction games to this case.

Let us start with the first possibility: Since each of these values has two possible truth values, 0 and 1, we can score probabilistic predictions by comparing them with these truth values for each of the possible values. This means in effect that we mimic a prediction game about a random variable with k values $v_1, \ldots, v_k$ by launching k prediction games about k binary events, v_m versus not-v_m, in parallel. The schema of this approach is depicted in figure 11.1.

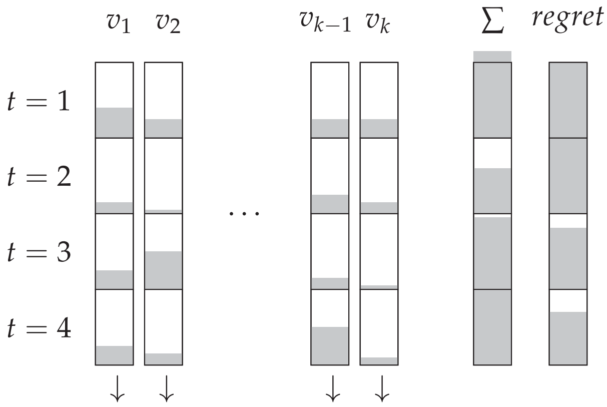

Figure 11.1: Example of launching k prediction games about single events parallel, one for each value of the value space. The bars under the labels of the values represent the probability predicted by the meta-inductive method; The bars representing the probability sum illustrates the possibility of incoherence (not always summing up to 1). Nevertheless, the forecast is optimal for each event value, which is indicated in the last column by guaranteed vanishing regrets for each predicted value w.r.t. the best candidate methods in the setting (regret is the difference between the per-round success rate of the candidate method and that of the meta-method).

We can define a measure for the predictive success regarding a value v_m as follows:

$$succ_{i,t}(v_m) = \frac{\sum_{u=1}^{t} 1 - \ell(Pr_{i,u}(v_m), val_u(v_m))}{t}$$

The decisive difference of this setting compared to the previous one is that now the success rates of the prediction methods are relative to elements of the value space: Each method has a success rate for each value v_m. Based on this we can define a weight $w_{i,t}(v_m)$ of method i for predicting event value v_m up to time t as follows:

$$w_{i,t}(v_m) = \frac{max(0, succ_{i,t}(v_m) - succ_{ami,t}(v_m))}{\sum_{j=1}^{n} max(0, succ_{j,t}(v_m) - succ_{ami,t}(v_m))}$$

Finally, based on these weights one might try to define a probabilistic meta-level method. E.g., for the method (AMI) one might define:

$$Pr_{ami,t+1}(v_m) = \sum_{i=1}^{n} w_{i,t}(v_m) \cdot Pr_{i,t+1}(v_m)$$

Of course we can now transfer the long run access optimality result of the foregoing section to such a meta-level method and prove that for each value v_m of Y's value space the meta-inductive prediction will approximate the maximum of the success rates of the predictors of v_m accessible in the setting. However, there is a problem: It can easily happen that there is not only one candidate method which is best at a given round for all values of the value space. In other words, the meta-inductive forecaster uses weights resulting from different prediction games which can lead to the result that its aggregated probabilities are incoherent. To see this, consider the following example:

- Let Y be a series of discrete random variables $Y_1, Y_2, \ldots$.
- $k = 3$, i.e. the value space consists of v_1, v_2, v_3.
- Let $n = 2$, i.e. the accessible candidate methods are Pr_1 and Pr_2. Now, let up to round u candidate method Pr_1 be a perfect expert in predicting v_1 and Pr_2 be a perfect expert in predicting v_2. Let up to round u Pr_1 completely fail regarding the predictions of v_2, v_3 and Pr_2 completely fail regarding predictions of v_1, v_3. Thus for all $t \leq u$: if $val_t(v_1) = 1$, then $Pr_{1,t}(v_1) = 1$ and $Pr_{2,t}(v_1) = 0$; and if $val_t(v_2) = 1$, then $Pr_{2,t}(v_2) = 1$ and $Pr_{1,t}(v_2) = 0$. Moreover if $val_t(v_3) = 1$ both fail, i.e. $Pr_{1,t}(v_3) = Pr_{2,t}(v_3) = 0$.
- So, the candidate predictions are such that their success rates at each round $t \leq u$ (for all convex loss functions without an additive term) are:

	$succ_{1,t}(v_i)$	$succ_{2,t}(v_i)$
v_1	100%	0%
v_2	0%	100%
v_3	0%	0%

- But then $Pr_{ami,u+1}(v_1) = Pr_{1,u+1}(v_1)$ and $Pr_{ami,u+1}(v_2) = Pr_{2,u+1}(v_2)$. Now assume that in round $u+1$ both of the candidate methods predict the value they were absolute experts up to round u, i.e. $Pr_{1,u+1}(v_1) = 1$ and $Pr_{2,u+1}(v_2) = 1$. Then the meta-inductive predictions are

  $$Pr_{ami,u+1}(v_1) = 1 \text{ and } Pr_{ami,u+1}(v_2) = 1$$

  which is probabilistically inconsistent.

So, although each individual provides a probabilistic forecast, pooling the forecasts according to this simple idea ends up with a forecast that is no longer probabilistically consistent. Regarding each value of the value space such a forecast is long run access optimal, however this optimality comes at cost of consistency.

One can try, of course, to restore consistency by normalising Pr_{ami}. Here the idea is to still calculate for each candidate method success rates that depend on the method's success regarding a specific value v_m of the value space. These success rates are then, in a second step, used for defining value-dependent weights for each candidate method. And these weights are again, in a third step, used to construct a prediction as above. However, additionally as a fourth step these predictions are normalised in order to guarantee probabilistic consistency:

$$Pr_{ami^*,t+1}(v_m) = \frac{Pr_{ami,t+1}(v_m)}{\sum\limits_{j=1}^{k} Pr_{ami,t+1}(v_j)}$$

A schema of such an implementation is illustrated in figure 11.2: Probabilistic forecasts consist no longer of parallel prediction games, but of combining parallel predictions by help of normalisation to a single probabilistic forecast.

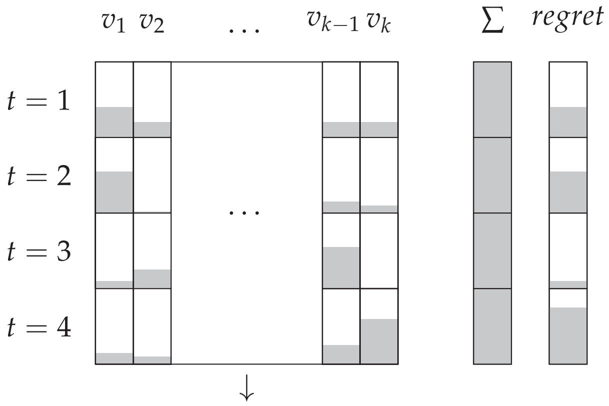

Figure 11.2: Example of a prediction game about single events, making a normalised prediction of the values of Y's value space. As in figure 11.1 above, the bars under the value labels represent the predicted probability. Note that for trivial reasons the predicted probabilities of the values sum up to 1 at each round. So they are probabilistically coherent. However, through normalisation there is no longer a guarantee for vanishing regret of the meta-inductive prediction w.r.t. each value of the value space. As indicated in the most right column, the regret might decrease and increase again. So we have probabilistically coherent, but suboptimal predictions.

However, such a meta-level method, although clearly probabilistically consistent, would no longer be long run access optimal. To see this, one just needs to specify the example from above and normalise the parallel meta-inductive predictions:

- Let us assume that we have three values v_1, v_2, v_3, two forecasters Pr_1, Pr_2 and for simplicity reasons let us assume that each of them gives at each round full probability to one of the values. Now, let us assume that the forecasts and the outcome are as follows:

t	1	2	3	4	5	6	7	8	...
Pr_1	$v_1 : 1.0$	$v_1 : 1.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 1.0$	$v_1 : 1.0$	$v_1 : 0.0$	$v_1 : 0.0$	...
	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	...
	$v_3 : 0.0$	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 1.0$	$v_3 : 0.0$	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 1.0$	...
Pr_2	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	...
	$v_2 : 1.0$	$v_2 : 0.0$	$v_2 : 1.0$	$v_2 : 0.0$	$v_2 : 1.0$	$v_2 : 0.0$	$v_2 : 1.0$	$v_2 : 0.0$	...
	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 0.0$	$v_3 : 1.0$	...
val	v_1	v_1	v_2	v_3	v_2	v_1	v_2	v_3	...

- Let us furthermore assume a linear loss function (similar counterexamples are possible with other convex loss functions). Then the success rates will converge to $succ_{1,t\to\infty}(v_1) = succ_{2,t\to\infty}(v_2) = 7/8$, $succ_{1,t\to\infty}(v_2) = succ_{2,t\to\infty}(v_1) = 5/8$, $succ_{1,t\to\infty}(v_3) = succ_{2,t\to\infty}(v_3) = 7/8$. Thus, after some point in time t^*, Pr_1 will gain full attractivity and weight in predicting v_1, Pr_2 full attractivity and weight in predicting v_2, and both get equal weight in predicting v_3. Hence, starting at $t^* + 1$ the unnormalised and the normalised predictions of the meta-level agents are:

t	t^*	$t^* + 1$	$t^* + 2$	$t^* + 3$	...
Pr_{ami}	$v_1 : 1.0$	$v_1 : 1.0$	$v_1 : 0.0$	$v_1 : 0.0$	...
	$v_2 : 1.0$	$v_2 : 0.0$	$v_2 : 1.0$	$v_2 : 0.0$	...
	$v_3 : 0.0$	$v_3 : 0.5$	$v_3 : 0.5$	$v_3 : 1.0$	...
Pr_{ami^*}	$v_1 : 0.5$	$v_1 : 0.\overline{66}$	$v_1 : 0.0$	$v_1 : 0.0$	...
	$v_2 : 0.5$	$v_2 : 0.0$	$v_2 : 0.\overline{66}$	$v_2 : 0.0$	...
	$v_3 : 0.0$	$v_3 : 0.\overline{33}$	$v_3 : 0.\overline{33}$	$v_3 : 1.0$	...
val	v_1/v_2	v_1	v_2	v_3	...

- But then—given, e.g., the natural loss function—the success rates of Pr_{ami^*} are: $succ_{ami^*}(v_1) = 19/24 < 7/8 = succ_1(v_1)$, $succ_{ami^*}(v_2) = 19/24 < 7/8 = succ_2(v_2)$, and $succ_{ami^*}(v_3) = 10/12 < 7/8 = succ_1(v_3) = succ_2(v_3)$. Hence, regarding all three values Pr_{ami^*} is long run *sub*optimal.

As things stand, a multiple parallel application of the meta-inductive framework to probability aggregation faces the dilemma of being either prone to *inconsistency* or *suboptimality*. We need to find a better way to apply meta-induction to probabilistic forecasting.

There is indeed a way to employ meta-inductive single event prediction game access optimality for probability aggregation in such a way that the meta-level method is both probabilistically consistent as well as access optimal. The crucial idea is to define a success measure for each method that is not relative to the values of Y's value space. In order to remain access optimal we will consider at each round the score of a method regarding *that value* which turned out to be the *true value* in the round. The schema of this approach is depicted in figure 11.3. This method of defining success is introduced in (Schurz 2019, sect.9.1); the same method of defining the success of a probabilistic forecaster is applied in sequential probability assignment (Cesa-Bianchi and Lugosi 2006, p. 248), but restricted to the logarithmic loss function. Here, in the context of strategies of probability aggregation, we introduce this method in a more general way applying to all convex loss functions.

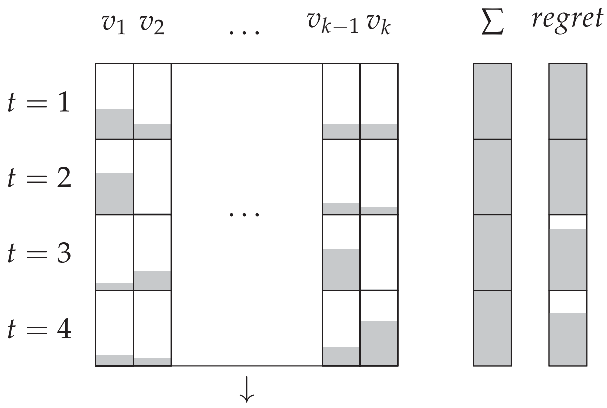

Figure 11.3: Example of a prediction game about single events using weights calculated out of predictions of those values which turned out to be true. Again, as in figure 11.1, the bars below the value labels represent the probability forecast. Note that at each round they sum up to 1 and that the probability forecast is optimal regarding the truth (the value which turned out to be the true one in each round), as indicated by the guaranteed vanishing regret. Hence we have a probabilistically coherent and optimal meta-inductive prediction method.

So, we define now a measure for the average per-round success in the probabilistic setting that is based on the loss of the prediction of that value which turned out to be the correct outcome in the round. We write this success rate as $\mathscr{s}ucc_{i,t}$:

$$\mathscr{s}ucc_{i,t} = \frac{\sum_{u=1}^{t} 1 - \ell(Pr_{i,u}(v_m^u), 1)}{t}$$
$$\text{where } v_m^u \text{ is that } v_m \text{ such that } val_u(v_m) = 1$$

Again we can define success-based weights, but this time without reference to a specific value of the value space:

$$w_{i,t} = \frac{max(0, \mathscr{s}ucc_{i,t} - \mathscr{s}ucc_{ami,t})}{\sum_{j=1}^{n} max(0, \mathscr{s}ucc_{j,t} - \mathscr{s}ucc_{ami,t})}$$

And again, by help of these weights we can define a meta-level probability aggregation function that aggregates the object-level probability functions by a success-based weighted arithmetic mean:

$$Pr_{\{1,\ldots,n\},t+1}(v_m) = Pr_{ami,t+1}(v_m) = \sum_{i=1}^{n} w_{i,t} \cdot Pr_{i,t+1}(v_m) \qquad (\text{AMI}^p)$$

This probability aggregation function is an instance of the meta-inductive method (AMI). For this reason, the long run optimality result regarding Pr_{mi} of (AMI) can be simply transferred to the probability aggregation rule $Pr_{\{1,\ldots,n\}} = Pr_{ami}$:

Theorem 11.1. *Given that ℓ is convex (where ℓ is used for determining $\mathscr{s}$ as defined above), then the forecaster Pr_{ami} (as defined in (AMI^p)) is long run access optimal:*

$$\lim_{t\to\infty} max(\mathscr{s}ucc_{1,t}, \ldots, \mathscr{s}ucc_{n,t}) - \mathscr{s}ucc_{ami,t} \quad \leq \quad 0$$

The same strategy can be applied also for proving optimality of the exponentially weighting meta-inductive probabilistic forecaster in accordance with (EAMI). The application is straightforward.

In conclusion, if probability aggregation is considered in a dynamical setting which allows one to compare the scores by calculating past success rates, then meta-inductive probability aggregation as presented here provides an epistemic rationale for using such success-based weights: It is simply because in doing so one has a guarantee for approaching or even outperforming the best predictive probabilities accessible in the setting.

Note that the only assumption in the optimality-result is that the used loss-function ℓ is convex. If, e.g., ℓ is the quadratic loss function, then the relative (per round) loss $1 - \mathscr{s}$ is equal to the normalised *Brier score* for binary events (see Brier 1950).

Up to now we have achieved an epistemic rationale for choosing weights used in linear probability aggregation in a success-based way. Let us now address the problem of providing an epistemic rationale for choosing weights used in geometric probability aggregation.

11.4 Meta-Inductive Geometric Probability Aggregation

The idea of the following approach to geometric probability aggregation and the proof of theorem 11.2 is thanks to, courtesy of, Schurz (see Feldbacher-Escamilla and Schurz manuscript).

We have seen in the preceding section that there is a way of aggregating probabilities by a linear success-based weighting rule which allows for long run access optimality. In this section we want to expand this result also to geometric success-based weighted probability aggregation. It is clear that according to (GM) there is no direct implementation of the meta-inductive optimality results of section 11.2 for geometrical rules, because these optimality results are only about linear success-based weighted predictions of single events. We succeeded already in transforming the optimality results from a setting of predictions about single events to the probabilistic case. Now we want to show how this result can be used further to allow also for proving optimality of a geometrical rule that uses success-based weights.

First, let us state what such a geometrical meta-level rule has to look like. In analogy to the instantiation of (AM) by the meta-level method (AMI^p), we aim at an instantiation of (GM) by a meta-level method:

$$Pr_{\{1,\ldots,n\},t+1}(v_m) = Pr_{gmi,t+1}(v_m) = \frac{1}{c} \cdot \prod_{i=1}^{n} Pr_{i,t+1}(v_m)^{gw_{i,t}}$$

(where $1/c$ is a factor needed for normalisation:

$$c = \sum_{j=1}^{k} \prod_{i=1}^{n} Pr_{i,t+1}(v_j)^{gw_{i,t}}$$

and the Pr_is are ϵ-regular, i.e. $Pr_i(v_m) \geq \epsilon > 0$;

for details regarding ϵ-regularity see equation (11.4) in the proof of theorem 11.2)

(GMI^p)

Second, in order to calculate weights that are success-based and allow for transferring an optimality result to such a meta-level rule, we want to highlight that the geometrical rule (GMI^p) can be re-stated as a linear rule similar to (AMI^p) aggregating logarithmic values:

$$\log(Pr_{gmi,t+1}(v_m)) = \sum_{i=1}^{n} gw_{i,t} \cdot \log(Pr_{i,t+1}(v_m)) - \log(c)$$

Third, the main idea of our implementation is to devise a prediction game which consists of logarithms of probabilistic forecasts. The weights of such a prediction game are success-based; then it is shown that they allow for applying the meta-inductive optimality result (AMI) as is done in (AMI^p), and finally, this result is transferred via the equation above to the geometrical rule (GMI^p) by fitting "geometrical" weights (gw_i) via "geometrical" scores and success rates (gsucc_i). The schema of this approach is provided in figure 11.4.

$$\begin{tikzcd}
Pr^*_{ami}, Pr^*_i \arrow[d, Rightarrow] & & Pr_{gmi}, Pr_i \arrow[ll, Rightarrow] \\
\mathit{succ}^*_{ami}, \mathit{succ}^*_i \arrow[d, Rightarrow] & & \mathit{gsucc}_{gmi}, \mathit{gsucc}_i \arrow[u, Rightarrow] \\
\sum_i w^*_i \cdot Pr^*_i \arrow[rr, phantom, "="] & & \prod_i Pr_i{}^{\mathit{gw}_i} \arrow[u, Rightarrow]
\end{tikzcd}$$

Figure 11.4: Schema of transferring the linear meta-inductive optimality result to the geometric aggregation rule. The *–variables are the variables of a logarithmic prediction game which is a certain instance of (AMI^p). For this instance the general meta-inductive optimality result holds, as was shown in section 11.3. One can equate this instance with (GMI^p). Now, via reverse engineering one can define success measures $\mathit{gsucc}_{gmi}, \mathit{gsucc}_i$ which allow for geometric meta-inductive optimality in the probabilistic prediction game (*–free variables).

Given such a procedure, an optimality result also holds for the geometric rule: We can devise a geometric scoring rule and based on it a definition for geometric success for the candidate methods as follows (see equation (11.7)) in the proof of theorem 11.2):

$$\mathit{gsucc}_{i,t} = \frac{1}{t} \cdot \log\left(\prod_{u=1}^{t} \frac{Pr_{i,u}(v^u_m)}{\epsilon} \right)$$

(where v^u_m is that v_m such that $val_u(v_m) = 1$,

and ϵ is the smallest real > 0 such that the Pr_is are ϵ-regular)

The weights for the candidate methods result from normalising their success rates:

$$\mathit{gw}_{i,t} = \frac{\max(0, \mathit{gsucc}_{i,t} - \mathit{gsucc}_{gmi,t})}{\sum_{j=1}^{n} \max(0, \mathit{gsucc}_{j,t} - \mathit{gsucc}_{gmi,t})}$$

The relative success rate $\mathit{gmisucc}_t$ of the geometric meta-inductive method Pr_{gmi} (GMI^p) is based on an instance of this success rate with $Pr_{gmi,u}(v^u_m)$, together with an additional factor c for "de-normalisation" (see equation (11.8) of the proof of theorem 11.2):

$$\mathit{gmisucc}_t = \frac{1}{t} \cdot \log\left(\prod_{u=1}^{t} \frac{Pr_{gmi,u}(v_m^u) \cdot c}{\epsilon}\right)$$

c is a specification of the normalisation term of (GM):

$$c = \sum_{j=1}^{k} \prod_{i=1}^{n} Pr_{i,u}(v_j)^{\mathit{gw}_{i,u}}$$

That the success measure for the candidate methods must differ from that of the geometric meta-inductivist results from the fact that geometric averaging of probabilities requires an additional step of re-normalising the resulting probability function; this step is not needed in their arithmetic averaging. Now, given this success measures it holds (the theorem and proof is due to Schurz—see Feldbacher-Escamilla and Schurz manuscript):

Theorem 11.2. *Pr_{gmi} as defined in* (GMI^p) *is long run access optimal w.r.t. the geometrical relative success measures gsucc and $\mathit{gmisucc}$:*

$$\lim_{t\to\infty} max(\mathit{gsucc}_{1,t}, \ldots, \mathit{gsucc}_{n,t}) - \mathit{gmisucc}_t \quad \leq \quad 0$$

Proof. Here we provide details for our approach to geometric meta-inductive probability aggregation: Recall the schema in figure 11.4. We go through it according to the following steps: We first define geometric pooling ①, then devise a game with predictions of logarithms of probabilities with an arithmetic meta-inductivist ②, define the respective success measures of this game ③, transform this game into a prediction game about probabilities with a geometric meta-inductivist ④, define—via backwards engineering—the respective success measures of this game ⑤, show that this is the success measure for geometric pooling and thus verify the optimality of the geometric meta-inductivist with respect to these success measures ⑥.

$$\begin{tikzcd}
 & \overset{\textcircled{2}}{Pr^*_{ami}, Pr^*_i} \arrow[d, Rightarrow] & \overset{\textcircled{1}}{Pr_{gmi}, Pr_i} \arrow[l, Rightarrow] & \\
\textcircled{3} & \mathit{succ}^*_{ami}, \mathit{succ}^*_i \arrow[d, Rightarrow] & \mathit{gsucc}_{gmi}, \mathit{gsucc}_i \arrow[u, Rightarrow, "\textcircled{6}"'] & \textcircled{5} \\
 & \sum_i w^*_i \cdot Pr^*_i \arrow[r, phantom, "=", "\textcircled{4}"'] & \prod_i Pr_i^{\mathit{gw}_i} \arrow[u, Rightarrow] &
\end{tikzcd}$$

①: We aim at the optimality of Pr_{gmi} as defined in (GMI^p).

Let us assume a prediction game whose task it is to predict the logarithm of the probabilities. We will put an *asterix* * over all variables of this prediction game. Then for all candidate methods $Pr_i \in \{Pr_1, \dots, Pr_n\}$, values of the value space $v_m \in \{v_1, \dots, v_k\}$, and for all rounds (time points) t it holds:

$$Pr^*_{i,t+1}(v_m) = \log(Pr_{i,t+1}(v_m)) \tag{11.1}$$

②: Now, applying probabilistic meta-induction to these candidate methods of the logarithmic prediction game yields, according to (AMI^p):

$$Pr^*_{ami,t+1}(v_m) = \sum_{i=1}^{n} w^*_{i,t} \cdot \underbrace{Pr^*_{i,t+1}(v_m)}_{\underset{((11.1))}{=} \log(Pr_{i,t+1}(v_m))} \tag{11.2}$$

Here the weights $w^*_{i,t}$ are success-based ($\mathit{succ}^*_{i,t}$) as follows:

$$w^*_{i,t} = \frac{max(0, \mathit{succ}^*_{i,t} - \mathit{succ}^*_{ami,t})}{\sum_{j=1}^{n} max(0, \mathit{succ}^*_{j,t} - \mathit{succ}^*_{ami,t})} \tag{11.3}$$

The relative (average per round) success $\mathit{succ}^*_{i,t}$ is based on a loss function. However, since these measures operate on predictions of the logarithm of probabilities, also the scores are logarithmic. In order to bound them, we assume that the probabilities of the candidate methods are ϵ-regular, i.e.:

> There is an $\epsilon > 0$ such that at any point in time t, for any probability or candidate method Pr_i, and all values v_m of the value space Val: $Pr_{i,t}(v_m) \geq \epsilon$
>
> (11.4)

Note that ϵ-regularity implies regularity as assumed in (GM), but not the other way round, so this assumption is stronger. By this assumption we know that $\mathit{succ}^*_{i,t}$ is bounded as follows: Since $Pr_{i,t+1}(v_m) \in [\epsilon, 1]$ we know that $Pr^*_{i,t+1}(v_m) \in [\log(\epsilon), 0]$. Hence, $\log(\epsilon)$ is the maximal logarithmic loss and $-\log(\epsilon) = \log(1/\epsilon)$ is the maximal logarithmic score. If we assume the natural loss function ($\ell(x,y) = |x,y|$), we get:

$$\mathit{succ}^*_{i,t} = \frac{\sum_{u=1}^{t} \log(1/\epsilon) + \log(Pr_{i,u}(v^u_m))}{t} \tag{11.5}$$

where v^u_m is that v_m such that $val_u(v_m) = 1$

Since (11.2) is an instantiation of (AMI^p) and we assumed a convex loss function, it follows from our investigation in section 11.3 that Pr^*_{ami} is long run access optimal. So we have defined the relevant success measures $\mathit{succ}^*_i, \mathit{succ}^*_{ami}$ for the logarithmic game ③✓.

We now transform this scoring measure to an ordinary prediction game whose task it is to predict the probabilities simpliciter. Re-transforming Pr_i^* to Pr_i is possible via $Pr_{i,t+1}(v_m) = e^{Pr_{i,t+1}^*(v_m)}$ (the inverse function of log). Similarly for the meta-inductive aggregation method:

$$Pr_{mi,t+1}(v_m) = e^{Pr_{ami,t+1}^*(v_m)} \underset{((11.2))}{=} \exp\left(\sum_{i=1}^{n} w_{i,t}^* \cdot \underbrace{Pr_{i,t+1}^*(v_m)}_{\underset{((11.1))}{=} \log(Pr_{i,t+1}(v_m))}\right) = \prod_{i=1}^{n} Pr_{i,t+1}(v_m)^{w_{i,t}^*} \tag{11.6}$$

Now, (11.6) resembles already (GMI^p), so ④✓. Only two things are different: First, the weights $w_{i,t}^*$ are still based on the logarithms of the predictions, and second, the normalisation factor $1/c$ (see (GMI^p) is missing.

Now we aim at defining a scoring measure which allows us to achieve long run optimality of the *normalised geometric aggregation* Pr_{gmi} of (GMI^p) compared to $Pr_1, \ldots, Pr_n$. The average per round score of the prediction game providing forecasts of the logarithms of probabilities can be used to construct a success measure which gets rid of the logarithm in the weights:

- We define the score $score_{i,t}$ of a candidate method Pr_i in round t as:

  $$score_{i,t} = Pr_{i,t}(v_m^t)/\epsilon$$

  where v_m^t is that value v_m of Val which turned out to be the true value at t. Each score is in the range $[1, 1/\epsilon]$. We can then define the *absolute geometric success* of a candidate method as the logarithm of the product of these scores ($\log \prod_{u=1}^{t} score_{i,u}$). Alternatively we could use a logarithmic loss function already for the one-round scores and define $\log(Pr_{i,t}(v_m^t)/\epsilon)$ as the score of one round. In this case the absolute success after t rounds would be given as the sum of these logarithmic scores. Both methods are equivalent. This is the reason why one finds both labels in the literature quite often used interchangeably: *geometric pooling* (due to the product) and *logarithmic pooling*. The *relative geometric success* is the *absolute geometric success* divided by t:

  $$gsucc_{i,t} = \frac{\log\left(\prod_{u=1}^{t} score_{i,u}\right)}{t} = \frac{\log\left(\prod_{u=1}^{t} \frac{Pr_{i,u}(v_m^u)}{\epsilon}\right)}{t} \tag{11.7}$$

- Note that from (11.5) we get: $succ_{i,t}^* = \frac{\sum_{u=1}^{t} \log(Pr_{i,u}(v_m^u)) - \log(\epsilon)}{t}$

- From this it follows: $succ_{i,t}^* = \frac{\log\left(\prod_{u=1}^{t} \frac{Pr_{i,u}(v_m^u)}{\epsilon}\right)}{t}$

- Since $\mathscr{succ}^*_{i,t} = \boldsymbol{\mathscr{gsucc}}_{i,t}$, we can also interpret $\mathscr{succ}^*_{i,t}$ as the *relative geometric success* of candidate method Pr_i up to round t.

- Up to now we have defined a success measure which allows us to get rid of the logarithms in the weights. However, this is not enough for the geometric meta-inductive method, since this method also uses normalisation. For this reason we have to define a success measure for this meta-inductivist which "de-normalises". We can do so by simply implementing the normalisation factor c into the score $score_{i,t}$:

$$gmiscore_t = c \cdot score_{gmi,t} = c \cdot \frac{Pr_{gmi,t}(v^t_m)}{\epsilon}$$

where $c = \sum_{j=1}^{k} \prod_{i=1}^{n} Pr_{i,t}(v_j)^{\boldsymbol{\mathscr{gw}}_{i,t}}$

and $\boldsymbol{\mathscr{gw}}_{i,t}$ is defined as usual, namely the normalisation of $\boldsymbol{\mathscr{gsucc}}_{i,t} = \mathscr{succ}_{i,t}$:

$$\boldsymbol{\mathscr{gw}}_{i,t} = \frac{\max(0, \boldsymbol{\mathscr{gsucc}}_{i,t} - \boldsymbol{\mathscr{gsucc}}_{gmi,t})}{\sum_{j=1}^{n} \max(0, \boldsymbol{\mathscr{gsucc}}_{j,t} - \boldsymbol{\mathscr{succ}}^*_{gmi,t})}$$

- Based on this "de-normalising" score we can define the *de-normalised relative geometric success* of Pr_{gmi} as:

$$\boldsymbol{\mathscr{gmis}}_t = \frac{1}{t} \cdot \log \left(\prod_{u=1}^{t} gmiscore_u \right) =$$

$$\frac{1}{t} \cdot \log \left(\prod_{u=1}^{t} \left(\underbrace{\frac{Pr_{gmi,u}(v^u_m)}{\epsilon} \cdot \underbrace{\sum_{j=1}^{k} \prod_{i=1}^{n} Pr_{i,u}(v_j)^{\boldsymbol{\mathscr{gw}}_{i,u}}}_{c}}_{gmiscore_u} \right) \right) \tag{11.8}$$

where $\boldsymbol{\mathscr{gw}}_{i,t} = \dfrac{\max(0, \boldsymbol{\mathscr{gsucc}}_{i,t} - \boldsymbol{\mathscr{gsucc}}_{gmi,t})}{\sum_{j=1}^{n} \max(0, \boldsymbol{\mathscr{gsucc}}_{j,t} - \boldsymbol{\mathscr{succ}}^*_{gmi,t})}$

So, we calculated the relevant success measures $\boldsymbol{\mathscr{gsucc}}_i, \boldsymbol{\mathscr{gsucc}}_{gmi}, \boldsymbol{\mathscr{gmisucc}}$ for the geometric game, hence ⑤✓

- We already know that the *relative geometric success* of a candidate method equals the relative success rate of the respective candidate method in the logarithmic game: $\mathscr{succ}^*_{i,t} = \boldsymbol{\mathscr{gsucc}}_{i,t}$

- We also know that the arithmetic meta-inductive method of the logarithmic game, i.e. Pr^*_{ami}, with its success rate succ^*_{ami} is long run access optimal.
- We can now show that the geometric meta-inductive method of the non-logarithmic game, i.e. Pr_{gmi}, with its success rate $\mathit{gmisucc}$ is also long run access optimal, by showing that for all rounds $t > 1$ (the special case of $t = 1$ does not matter, because we can simply stipulate equivalence there): $\mathit{succ}^*_{ami,t} = \mathit{gmisucc}_t$:
  - According to our backwards engineering above, the relative geometric score of the meta-inductive method is as given in (11.8):

$$\mathit{gmisucc}_t = \frac{1}{t} \cdot \log \prod_{u=1}^{t} \left(Pr_{gmi,u}(v_m^u)/\epsilon \cdot c\right)$$

  - Now, according to (GMI^p):

$$Pr_{gmi,u}(v_m^u) = \frac{\prod_{i=1}^{n}(Pr_{i,u}(v_m^u)^{gw_{i,u}})}{c}$$

    so, the normalisation factor c cancels and we get:

$$\mathit{gmisucc}_t = \frac{1}{t} \cdot \log \prod_{u=1}^{t} \left(\frac{\prod_{i=1}^{n}(Pr_{i,u}(v_m^u)^{gw_{i,u}})}{\epsilon} \right)$$

  - By reformulation we get:

$$\mathit{gmisucc}_t = \frac{1}{t} \cdot \sum_{u=1}^{t} \left(\sum_{i=1}^{n} \left(gw_{i,u} \cdot \log(Pr_{i,u}(v_m^u))\right) - \log(\epsilon) \right)$$

  - Since $\sum_{i=1}^{n} gw_{i,u} = 1$ we get:

$$\mathit{gmisucc}_t = \frac{1}{t} \cdot \sum_{u=1}^{t} \left(\sum_{i=1}^{n} \left(gw_{i,u} \cdot (\underbrace{\log(Pr_{i,u}(v_m^u))}_{\underset{(11.1)}{=} Pr^*_{i,u}(v_m^u)} \underbrace{- \log(\epsilon))}_{=\log(1/\epsilon)} \right) \right) = \mathit{succ}^*_{ami,t}$$

This completes the proof: ⑥✓.

We conjecture that this result can be generalised by using more appropriate loss functions different from the natural one. In particular we conjecture that by this, one could get rid of the "de-normalisation" in the success rate of Pr_{gmi}. However this a very complex topic and work for future research. □

This result also shows that geometric probability aggregation can be performed in a success-based way such that long run access optimality of such an aggregation can be guaranteed. This also provides an epistemic rationale for geometric aggregation. Note, however, that due to the restrictions of geometrical pooling this result is much less general. Whereas linear pooling allows for defining a variety of success-based weights which prove to guarantee long run access optimality in case the underlying loss function is convex, for geometrical pooling we were only able to show that there exists at least a success-based weighting method that proves to guarantee optimality in the long run.

We can sum up the main findings of this chapter as follows: We have argued for a new solution to the problem of weighted probability aggregation. We have seen that some general constraints determine families of aggregation rules. However, even if arguments can be put forward for deciding for a particular family, in the classical approach the choice of an exact aggregation rule of the respective family remains epistemically undetermined. We have argued that a success-based calculation of weights—as is done in the framework of meta-induction—allows for a much more precise choice. Success-based weighting also provides a rationale for such a choice, since it guarantees long run optimality in probabilistic prediction tasks. Whereas the exact choice of the weights for linear or geometric probability aggregation might still depend on the context and purposes in question, all such choices can be epistemically justified as long as the respective conditions of the optimality results are justified.

Chapter 12

The Wisdom of the Crowds

This chapter briefly sketches the historical discussion and first analysis of a wise crowd effect, and provides a general characterisation. Afterwards, Condorcet's wise crowd effects in probabilistic predictions are outlined. Subsequently, wise crowd effects in non-probabilistic predictions and their underlying assumptions are investigated. Finally, it is shown how meta-induction can be interpreted as cashing out a wise crowd effect.

Oftentimes individuals have to act together in order to achieve their individual, a shared or a common goal. This holds true not only for the practical realm of (practical) decision making, but also for the epistemic realm. Science is a collective endeavour in the sense that all aspects of scientific investigation—the discovery, the justification, the utilisation of theories—depend on collective action, may it be collective belief formation or collective decision making. Logical positivists and empiricists have often stressed the importance of such collective action also for philosophy. So, e.g., Carnap wrote in his first influential work, the *Aufbau*, the following (in a similar vein was Neurath's discussion of expert knowledge in section 9.4):

> "The new type of philosophy has arisen in close contact with the work of the special sciences, especially mathematics and physics. Consequently they have taken the strict and responsible orientation of the scientific investigator as their guideline for philosophical work, while the attitude of the traditional philosopher is more like that of a poet. This new attitude not only changes the style of thinking but also the type of problem that is posed. The individual no longer undertakes to erect in one bold stroke an entire system of philosophy. Rather, each works at his special place within the one unified science." (Carnap 1928/2003, p.xvi)

Now, in evaluating collective action, one can take individual or other collective action as a benchmark. A quite common way of evaluation is to compare the performance of a collective with respect to the performance of an average individual of that collective. The spectrum of outcomes of such an evaluation is quite broad. It ranges from the claim that collectives are *mad*: “Men, it has been well said, think in herds; it will be seen that they go mad in herds, while they only recover their senses slowly, and one by one” (Mackay 1852, p.viii); to the claim that collectives are *wise* and *knowledgeable*: “We are more likely to attribute [...] success to a few smart people in the crowd than to the crowd itself. [... However] chasing the expert is a mistake, and a costly one at that. We should stop hunting and ask the crowd [...] instead. Chances are, it knows” (Surowiecki 2005, p.xv). Clearly, this is about extreme cases of collective performance. More typical is, however, some form of intermediary performance: Collective action is often also considered as a trade-off where individuals almost literally trade with inferiorities and merits, and gain advantages in one area at the cost of disadvantages in another.

A short note on terminology is in place: Here we speak of ‘collective action’ and attribute several epistemic attitudes to a collective as if such a collective were a single epistemic agent. There are many similarities between single individuals and collectives which seem to grant such a terminology (so, e.g., List and Pettit 2011, argue prominently for group agency by help of similarities). However, clearly, there are also many dissimilarities between single individuals and collectives for which reason one ought to be cautious when using such a terminology (collectives do not have consciousness in the ordinary sense, etc.). Our choice of terminology is not intended to make any deep ontological, action theoretical, and moral assumptions. We think that the agency way of speaking is convenient, but we agree also that a rephrasing of the claims made here in terms of *collective effects*, etc. is completely fine too.

In this chapter we deal with characterisations of cases where collective action performs well in the sense of approaching wisdom or knowledge. In section 12.1 we discuss the most famous example of a wise crowd and provide a very general characterisation. In section 12.2 we briefly discuss the most famous theorem on wise crowd characteristics, namely the so-called *Condorcet jury theorem*. In section 12.3 we investigate another result on wise crowd characteristics in a setting which is closer to our prediction setting, namely the so-called *The Crowd Beats the Average Law*. Finally, in section 12.4 we hint at an interpretation of meta-inductive optimality as another result of such wise crowd characteristics.

12.1 Wise Crowd Effects

Let us begin with a true story about the first influential empirical observation and theoretical analysis of a wise crowd effect. The main proponent of this story is Sir Francis Galton (1822-1911) who was a highly versatile scientist, a real polymath (for the following see Gillham 2001). Galton invented and introduced statistical methods and concepts like *correlation* and *regression towards the mean*. He founded several disciplines as, e.g., psychometrics. He devised the methodology for creating weather maps, and also explored Africa. It is interesting to note that he was a cousin of Charles Darwin (1809-1882) and also highly interested in heredity. Due to this he also became the founder of *eugenics*, a discipline concerned with topics which, back at his time, were simply considered to be subject to ordinary scientific *Inquiries into Human Faculty and Its Development* (1883). But what is most relevant for us here is that Galton became also famous for uncovering and investigating an impressive wise crowd effect (see Surowiecki 2005, pp.xiff): Jack of all trades as he was, his interest in inheritance lead Galton also to an interest in livestock. So, one day in fall 1906 he headed to a cattle show at Plymouth in order to get some more impressions on results of England's farmers' breeding capabilities. By accident he recognised that at the fair also an ox-weight-judging competition was going on, and this immediately called the eugenicist and statistician in him.

As an eugenicist he thought that intelligence and intellectual abilities are much more influenced by nature than *nurture* which are influences after conception as, e.g., social influences. Already in his *Hereditary Genius. An inquiry into its laws and consequences* (1869) he argued with help of historiometrical means that genius is distributed in clusters around family lineages and not widespread among the whole population. For this purpose he studied the family trees of influential people in society (judges, politicians, important people of the military), art (poets, musicians, painters), and science (see Galton 1869). And he interpreted this clustering result as an indicator for the hereditary of genius. Taking in the stance of an engineer, he considered this as a reason to suggest intellectual enhancement or at least to avoid thinning of intellectual abilities in society. These were the main aims of eugenics. The former aim was more due to evolutionary optimism regarding human species and new possibilities one might open up by this. Whereas the latter aim was more due to pessimism about the increased influence of the general population which was considered to be intellectually inferior to formerly more powerful classes.

Now, given this background, Galton thought that such a weight-judging competition allows for a further study of the distribution of ability (this time not necessarily intellectual ones) in the general population. Since Galton was perhaps one of the most rigid adherents of the Pythagorean programme (*All things are number*—he expressed this in his motto "*Whenever you can, count.*" (see Pearson 1924, p.340)), he must have felt really lucky to have come across such an opportunity for a quantitative analysis. This the more since, as he described it, the setting of this competition seemed to be almost perfect for such a study:

> "A weight-judging competition was carried on at the annual show of the West of England Fat Stock and Poultry Exhibition recently held at Plymouth. A fat ox having been selected, competitors bought stamped and numbered cards, for 6*d*. each, on which to inscribe their respective names, addresses, and estimates of what the ox would weigh after it had been slaughtered and "dressed." Those who guessed most successfully received prizes. About 800 tickets were issued, which were kindly lent me for examination after they had fulfilled their immediate purpose. These afforded excellent material. The judgments were unbiassed by passion and uninfluenced by oratory and the like. The sixpenny fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best. The competitors included butchers and farmers, some of whom were highly expert in judging the weight of cattle; others were probably guided by such information as they might pick up, and by their own fancies." (Galton 1907c, pp.450f)

Without further ado he borrowed the tickets from the organisers of the competition and ran statistical tests on the numbers of the weight guessing competition. Different to his former study on genius as mentioned above, he was not only interested in the distribution of the most accurate judgements or predictions, but also in the accuracy of the judgement or prediction of the *average* individual.

Galton was quite clear on what he considered to be *the* average individual; for this purpose he even had introduced a new notion to statistics in anterior writings: It was the *median* Galton considered to be the relevant measure:

> "How can the right conclusion be reached, considering that there may be as many different estimates as there are members? That conclusion is clearly *not* the *average* of all the estimates, which would give a voting power to "cranks" in proportion to their crankiness. One absurdly large or small estimate would leave a greater impress on the result than one of reasonable amount, and the more an estimate diverges from the bulk of the rest, the more influence would it exert. I wish to point out that the estimate to which least objection can be raised is the *middlemost* estimate, the number of votes that it is too high being exactly balanced by the number of votes that it is too low. Every other estimate is condemned by a majority of voters as being either too high or too low, the middlemost alone escaping this condemnation." (Galton 1907a, p.414)
>
> And:

> "I endeavoured in the memoirs just mentioned, to show the appropriateness of utilising the *Median* vote in Councils and in Juries, whenever they have to consider money questions. Each juryman has his own view of what the sum should be. I will suppose each of them to be written down. The best interpretation of their collective view is to my mind *certainly not* the average, because the wider the deviation of an individual member from the average of the rest, the more largely would it effect the result. In short, unwisdom is given greater weight than wisdom." (Galton 1908, p.281)

Now, Galton was quite impressed that in case of the ox-weight-judgement competition the mean was a very good predictor:

> "According to the democratic principle of "one vote one value," the middlemost estimate expresses the *vox populi*, every other estimate being condemned as too low or too high by a majority of the voters [...]. Now the middlemost estimate is 1207 lb., and the weight of the dressed ox proved to be 1198 lb.; so the vox populi was in this case 9 lb., or 0.8 per cent. of the whole weight too high." (Galton 1907c, p.451)

This led Galton, a former pessimist with regards to the influence of an empowered general population in the reasonableness of collective judgements, to the more optimistic conclusion: "This result is, I think, more creditable to the trustworthiness of a democratic judgment than might have been expected" (Galton 1907c, p.451). Galton published his finding 1907 in *Nature*. A brief discussion arose and people were also interested in the average prediction. In (Galton 1907b) he reported that the arithmetic *mean* of the almost 800 estimations was 1197 pounds, i.e. just 1 pound deviating from the true value. So, regardless of whether one considers the median or the mean as relevant for the *average* individual, the guessing crowd was almost a perfect predictor. What are the reasons for this almost perfect performance of the crowd? Galton himself did not analyse the case further. However, in his description of the case and the preceding discussion he refers to one property which is nowadays considered to be the main condition relevant for a wise crowd effect to show up, namely *diversity* (see Page 2007). Recall, Galton described the setting of the ox-weight-judgement competition as almost perfect inasmuch as "the judgments were *unbiassed* [...] and *uninfluenced* by oratory and the like" (see Galton 1907c, p.450). And also in describing how an optimal jury decision is achieved, he states as a condition "for each juryman to write his *own independent* estimate on a separate slip of paper" (Galton 1907a, p.414). We will discuss how to exactly spell out these conditions in the subsequent sections.

Now, the case presented by Galton is a case where the "*vox populi*" was impressively close to the true outcome. However, it seems that accuracy of the "*vox populi*" in more relative terms suffices already for reducing prejudices against collective judgements—so, e.g., if the "*vox populi*" were not that accurate, Galton might have been still impressed if it had outperformed the experts' judgements. For this reason, in general one speaks of a *wise crowd effect*, if a collective performs better than an individual or an other part of the collective does. As we have mentioned at the beginning of this part of the book, the problem of the wisdom of the crowds concerns the question under which conditions which forms of aggregation methods produce such a wise crowd effect. In this chapter we study different such conditions and aggregation methods.

Several specifications and remarks of this general characterisation of a wise crowd effect are in place. The relevant keywords are *wise*, *better performance*, and the relatum for comparison, namely *individual or other part of the collective*. Firstly, *wise* is to be understood very broadly. As used here, the notion is not linked to some form of deeper understanding or understanding in general. Rather, it is just used for one component of the much more general notion, namely the feature that wise actions or decisions are typically more often correct than others. Furthermore, usually a wise action is considered to fulfil high standards in absolute terms. Again, the notion we employ here is much weaker and demands satisfying high standards in relative terms only. To illustrate this, consider as an analogy the relation between *truth* and *being closer to the truth than*: One of the most ambitious epistemic aims is it to end up with informative theories that are true (see, e.g., our elaboration on *absolute learnability* in section 5.3); however, often there are no epistemic means to achieve this end for which reason we aim at informative theories that are closer to the truth than their competitors (see our elaboration on *relative learnability* in section 5.4). Similarly with the notion of *wisdom* used here: An action is called 'wise', if it is better compared to its alternatives; clearly, it would be advantageous, if it were also a *good* action; however, this is no necessary condition, if none of the competing actions is good. This brings us directly to the next notion, namely *better performance*.

Secondly, '*better performance*' means to provide better predictions. Predictions can be non-probabilistic or probabilistic. A non-probabilistic prediction is better than another one, if it is more accurate than the other one in terms of closer to the true value. A probabilistic prediction is better than another one, if it predicts the true event with higher probability than the other prediction does. In section 12.2 we will consider the probabilistic case; in section 12.3 the non-probabilistic one.

Thirdly, let us specify what is meant with '*individual or other part of the collective*': Usually, a collective consists of quite heterogeneous individuals. So, comparing a collective action with that of different individuals usually leads to different results. It is quite natural to use the *average* individual of the collective as a benchmark. Later on, when we discuss non-probabilistic predictions in detail, we will concentrate on this case. However, in principle one could also use as a benchmark other individuals than *the* average individual. So, e.g., one could also take as a benchmark the worst performing individuals and put forward the constraint that the collective should perform better or at least not worse than these individuals do, i.e. that the collective should be not strongly suboptimal, in order to be wise. In the light of the *madness* evaluation of collective action, it seems reasonable to assume that Mackay (1852) would have been fine with such a low benchmark: Any aggregation mechanism which prevents that individuals "go mad in herds" should be reasonable. Also our argument for using a *convex* loss function in online regression (section 3.4) was based on such a low benchmark only: If one demands of a learning algorithm which just averages predictions (see definition 3.31) that it should not be strongly suboptimal, i.e. outperformed even by the worst individuals, then employing a convex loss function is a suitable epistemic means to achieve this end (see theorem 3.32).

Another possibility is do not use the *average* individual as a benchmark, and also not "the weakest link" in the chain clinging together the collective, but to use the best individuals (the "strongest links") of the collective as a benchmark. Characterising conditions for a wise crowd effect based on such a high benchmark is quite demanding. As we will argue in section 12.4, if one expands the horizon from single predictions to long run prediction series, then the theory of meta-induction characterises conditions and aggregation methods for such wise crowd effects.

There is also the possibility to use another collective as a benchmark. So, e.g., one might wonder whether, and if so, under which conditions a collective performs better than one of its sub-collectives. Figuring out conditions and aggregation methods for such a wise crowd effect is relevant for the question of how to design collectives: When is it better to increase the size of a collective etc.? The investigation in section 12.2 is also about wise crowd effects with sub-collectives as a benchmark.

Finally, combinatorially speaking there is even a further possibility to characterise wise crowd effects via different relevant relata: We have spoken about comparing a collective with an individual and comparing a collective with a sub-collective. Now, it is interesting to note that conditions and aggregation methods for wise crowd effects can be also defined for the borderline case where the collective equals an individual. Hinting at this results concern us in the remainder of this section.

Recent psychological studies show that wise crowd effects arise also in case one averages over the judgements, estimations or predictions of one and the same individual. This phenomenon is called the *crowd within effect*. E.g. Vul and Pashler (2008) have performed a study where people were asked to answer questions probing their real-world knowledge as, e.g., estimating the percentage of the world's airports in the United States. Half the participants had to make a first guess immediately followed by a second guess. The others had to make a first guess and a second guess three weeks later (none of them knew when making the first guess that they are expected to make a second guess also). Now, Vul and Pashler (see 2008, p.646) found out that in both subgroups the average of both guesses were more accurate than each single guess. Furthermore, the averages in the group with three weeks delay between the guesses were significantly more accurate than those of the other group. Given that diversity is a main condition relevant for wise crowd effects, the three weeks delay might be interpreted as producing further diversity in the individual's guesses, and hence explain the higher accuracy. The main result of this study is depicted in figure 12.1.

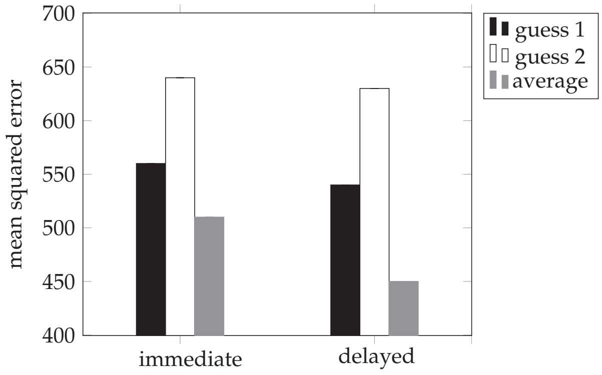

Figure 12.1: Results of the crowd within experiment of Vul and Pashler (see 2008, p.646): The study was performed on 428 participants; each of them were asked 8 questions; the "immediate" group (214 members) had to provide a second guess after their first guess with their first guess in front of them; the "delayed" group (214 members) had to provide a second guess after 3 weeks.

Herzog and Hertwig (2009) have shown another wise crowd effect with different means for achieving diversity, namely by what they call *dialectical bootstrapping*: Dialectical bootstrapping asks the participants to rethink their estimations along the following line:

> "First, assume that your first estimate is off the mark. Second, think about a few reasons why that could be. Which assumptions and considerations could have been wrong? Third, what do these new considerations imply? Was the first estimate rather too high or too low? Fourth, based on this new perspective, make a second, alternative estimate." (Herzog and Hertwig 2009, p.234)

The task in their study with 101 participants was to provide estimations of the year of 40 generally familiar historical events (16th–19th century, 10 questions each). It turned out that the first estimate was on average 130.8 years off, whereas dialectical bootstrapping was on average only 123.2 years off.

Interestingly, dialectical bootstrapping seems to be not just a curious phenomenon that can be simply subordinated to the class of wise crowd effects, rather, it seems to become also more theoretically interesting by allowing further systematisation of formal decision theory. So, e.g., in a recent paper Hartmann (2017) applies dialectical bootstrapping for providing a rationale for *prospect theory* of Kahneman and Tversky (1979).

So much for the first class of wise crowd effects, namely the case where an individual fares better, if it averages *diverse educated guesses* of her. Now, let us come to the other cases. In the next section we start with the case of wise crowd effects in probabilistic predictions.

12.2 Condorcet Juries

In the preceding section we have claimed that Galton was one of the first to discover and formally investigate a wise crowd effect in detail. However, this claim needs modification, if we look back further in history and take into account the mathematical investigations and arguments of French mathematicians, philosophers, scientists, and politicians of the 18nth and 19nth century.

The time we are speaking about centres around the *Enlightenment* (beginning with the death of *Louis XIV*, 1715 until 1789) and the *French Revolution* (beginning with the storming of the Bastille 1789 until Napoleon's first *coup* in 1799), where French monarchy was overthrown, and republic established. This political climate brought it with it, that the problem of aggregating judgements gained practically more weight; roughly speaking: In absolute monarchy or dictatorship it is all about one single judgement (*L'État c'est moi!*), whereas in republic and democracy it is about all the individual judgements that need to be condensed to some single ones in order to be dealt with. As is often the case, what gains practically more weight does so also in theory, for which reason *electoral studies* boomed. The first main works applying mathematical methodology to study social decision (i.e. the forerunners of social choice theory) were published around the 1780s. And its main proponents were also important figures of the Enlightenment. As D. Black (1986, p.183) describes it:

> "The second half of the eighteenth century in France was one of the outstanding epochs of scientific thought. Science had felt its strength and its impulse and did not know what barriers it might not cross. The hope had sprung up to carry the methods of rigorous and mathematical thought beyond the physical and into the realms of the human sciences."

Now, the three main figures of early electoral studies mentioned by D. Black (see 1986, pp.156ff) are Jean-Charles Chevalier de Borda (1733-1799), Nicolas, Marquis de Condorcet (1743-1794), and Pierre-Simon, Marquis de Laplace (1749-1827). Borda argued for a method known today as *Borda count*, where individuals rank alternatives according to their preferences. The ranks are mapped to points, where highest ranked alternatives, i.e. the most preferred ones, gain the highest number of points. And the aggregated judgement consists of that alternative(s) which gain highest points in summing over all individuals. Laplace worked out in detail the probabilistic arguments underlying Borda count and other forms of aggregation. However, out of this triumvirate Borda-Condorcet-Laplace, it is Condorcet who seems to be most entrenched with electoral studies: "Borda and Laplace discuss the matter no further than is necessary to establish that one form of election is good and the others, *ipso facto*, defective. Condorcet set out with a wider end in view [... which] gives a far more vital account of the nature of elections and of group decision-taking than any other" (see D. Black 1986, pp.184f). As we will see now, Condorcet's investigations are also highly relevant for characterising wise crowd effects.

Perhaps most famous is the so-called *Condorcet jury theorem* which was proven and published as early as 1785 by Condorcet. It states that majority aggregation produces an (estimated) wise crowd effect in deciding between two alternatives under the condition that the individual opinions are more likely to be true than false and that they are independent. More specifically, given these conditions, the probability that majority aggregation is correct increases as the number of individuals satisfying these conditions increases, and approaches 1 in case the number of individuals is infinite (see Condorcet 1785).

The reception of Condorcet's result underwent a remarkable development. Condorcet stated it in 1785 in his *Essay on the Application of Analysis to the Probability of Majority Decisions* (for a modern reconstruction see Courgeau 2012, pp.116ff). A rigorous formal proof of a more general result was provided in Laplace's *Analytic Theory of Probabilities* of 1812. In the 19th century Condorcet's *Essay* was not very well received and seemed to have gone forgotten. The main tenor of his critiques seem to be that the *programme* of Daniel Bernoulli (1700-1782) to apply the theory of probability to the civil, moral and economic realm, was taken up in the *Essay*, but that Condorcet was more "enthusiastic rather than scientifically exact", i.e. that he was mathematically wrong. D. Black (1986, p.184f) argues that this negative interpretation of Condorcet is mainly due to issues regarding notation: "symbolism [of the *Essay*] purported to belong to the mathematical theory of probability, but would now be regarded as a primitive topology, though this branch has only been developed quite recently; [... for this reason Condorcet's theory] was buried away at various parts of a long and difficult book and its meaning must be wrung from Condorcet's stilted and crabbed sentences." The first edition from 1958 of D. Black (1986) is also considered to be the source which rediscovered the theorem and its relevance for electoral studies (see Dietrich and Spiekermann 2013, p.88).

In what follows we describe a more general wise crowd effect based on modern probability theory and show then how *a* wise crowd effect described by Condorcet can be embedded as a special instance. We have mentioned already that Daniel Bernoulli had in mind a research programme of applying probability theory to the social realm. However, it is a theorem of Daniel's uncle, Jacob Bernoulli (1655-1705), "discoverer of probability theory and the Bernoulli numbers", which we are going to employ here, namely the *weak law of large numbers* (see Howson and Urbach 2006, sect.2.m).

As described in section 2.1, according to the weak law of large numbers, the average of the results of a large number of independent and identically distributed (i.i.d.) random experiments will be close to the expected value of the random experiment and will tend to become closer, the more such experiments are performed. Formally: Given a series of random variables (experiments) $Z_1, Z_2, \ldots$ with value space $\{z_1, \ldots, z_k\}$ which are independently and identically distributed (i.i.d.) around the mean z:

- Independence: $Pr(Z_i = z | Z_j = z) = Pr(Z_i = z)$ (for all $z \in \{z_1, \ldots, z_k\}, i \neq j \in \mathbb{N}$)
- Identity: $\mathbb{E}[Z_i] = z = Pr(Z_i = z_1) \cdot z_1 + \cdots + Pr(Z_i = z_k) \cdot z_k$ (for all $i \in \mathbb{N}$)

It holds that (where we operate on the outcomes of $Z_1, Z_2, \ldots$):

$$\text{For all } \varepsilon > 0 : \lim_{n \to \infty} Pr\left(\left| \frac{Z_1 + \cdots + Z_n}{n} - z \right| < \varepsilon \right) = 1$$

So, the law states that given *independence* and *identity*, averaging among the outcomes of the random experiments comes arbitrarily close to the expected value, once the number of such random experiments becomes arbitrarily high.

Let us quickly illustrate this by help of an example: Assume $Z_1, Z_2, \ldots$ to be random variables representing independently rolling a "fair" dice. Then $\mathbb{E}[Z_i] = \frac{1}{6} \cdot 1 + \cdots + \frac{1}{6} \cdot 6 = 3.5$. Now, the law of large numbers states that with an increasing number of rolling dies, their average tends towards the expected value, i.e. the average of the outcomes of $Z_1, \ldots, Z_n$ ($(Z_1 + \cdots + Z_n)/n$) tends better towards 3.5 than $Z_1, \ldots, Z_m$ ($(Z_1 + \cdots + Z_m)/m$), if $n \gg m$.

We show now, how this can be employed for characterising a wise crowd effect for probabilistic predictions (that Condorcet's theorem follows from the law of large numbers is a well-known fact in the literature; see, e.g., Dietrich 2008, appendix A): Assume a set $\{Z_1, Z_2, \ldots\}$ of probabilistic predictions of the value $z \in [0,1]$ (z is one of $\{z_1, \ldots, z_k\}$) with *identical* expected values $\mathbb{E}[Z_1] = \mathbb{E}[Z_2] = \cdots = z$. Assume furthermore that the predictors $Z_1, Z_2, \ldots$ are probabilistically independent. And furthermore, as indicated already by our notation, assume that the set is infinite. Let us assume that this infinite crowd aggregates by averaging. We denote the prediction of the value z of the infinite crowd by 'Z_{av}'. So $Z_{av} = \lim\limits_{n\to\infty} \left(\frac{Z_1+\cdots+Z_n}{n}\right)$. Now, we can define the expected loss of an individual prediction by an absolute loss function ℓ as $\mathbb{E}[\ell_i] = \ell(\mathbb{E}[Z_i], z)$, but since we assumed that $\mathbb{E}[Z_i] = z$ we automatically get $\mathbb{E}[\ell_i] = 0$. What is more relevant is the *actual* loss of an individual prediction, which is:

$$\ell_i = \ell(Z_i, z)$$

Let us assume that in fact all individuals have a positive loss which is the same across all individuals, i.e. $\ell_i = \ell_j > 0$ (for all $i, j \in \mathbb{N}$). Since they all have the same positive loss, each individual prediction represents also the predictive performance of the *average* individual. Let us denote the latter by $\ell_{\varnothing\{1,2,\ldots\}} = \ell_1 = \ell_2 = \cdots$. Since by assumption the average individual has a positive loss, it holds:

$$\ell_{\varnothing\{1,2,\ldots\}} = \varepsilon_1 > 0 \text{ (for some } \varepsilon_1\text{)}$$

Now, by the law of large numbers we get: For all $\varepsilon > 0$: $\lim\limits_{n\to\infty} Pr\left(\ell_{av(n)} < \varepsilon\right) = 1$ (where ℓ is the absolute loss function, i.e.: $\ell_{av(n)} = \ell(Z_{av(n)}, z) = |Z_{av(n)} - z|$). Hence, we know that $\ell_{av(n)}$ is also strictly smaller than ε_1, and hence:

$$\lim_{n\to\infty} Pr\left(\ell_{av(n)} < \ell_{\varnothing\{1,2,\ldots\}}\right) = 1$$

So, we have a probabilistic wise crowd effect stating that an infinite collective of individuals who make not perfect predictions *almost certainly* outperforms the average individual. If we use s as the inverse of ℓ within $[0,1]$, then we can also say:

$$\lim_{n\to\infty} s_{av(n)} > s_{\varnothing\{1,2,\ldots\}} \text{ almost certainly (i.e. with } Pr \text{ 1)}$$

We can now embed one part of Condorcet's jury theorem in this framework. The basic theme of his *Essai* from 1785 concerns the probability of a collective to take a correct decision. Recall, according to our summary in the introduction Condorcet's jury theorem states that if the votes of a jury are independent and (equally) competent, then majorities are more likely to select the correct opinion, i.e. the probability that a majority selects the correct opinion is greater than the probability of the individual having a correct opinion. This probability increases with the number of voters and approaches 1 if the number of voters gets arbitrarily high. Here are some key passages of the *Essai* (see Condorcet 1785):

“Nous supposerons d’abord que tous ceux qui donnent leurs voix, ont une égale sagacité, une égale justesse d’esprit dont ils ont fait également usage, qu’ils font tous animés d’un égal esprit de justice, enfin que chacun d’eux a voté d’après lui-même, comme il arriveroit fi chacun prononçoit séparément son avis[.]” (p.3)

[Independence]: We shall first assume that all those who give their voices have the same sagacity and equanimity of mind, which they have also used, that they all have an equal spirit of justice, and that each of them *voted according to themselves*, as it would happen if *each one pronounces separately her opinion*.

“Nous supposerons en général que v représente le nombre de fois que l’opinion d’un des Votans doit être conforme à la vérité, & e le nombre de fois qu’elle doit être contraireà la vérité fur un nombre $v + e$ de décisions; & pour abréger, nous supposerons $v + e = 1$ en général.” (p.3)
“Ainsi, par exemple, [...] $v > e$[.]” (p.6)

[Equal competence]: We will usually assume that v represents the number of times that the opinion of one of the Votans conforms to the truth, and e the number of times it is contrary to the truth for a number of $v + e$ decisions; to abbreviate, we will assume $v + e = 1$ in general.
For example: $v > e$.

“Le nombre des Votans est $2q + 1$, & l’on cherche la probabilité de la pluralité d’une seule voix.” (p.3)
“[L]orsque $v > e$, la probabilité pour que la décision soit conforme à la vérité, augmentera fans cesse, en augmentant le nombre des Votans;” (p.6)
“[La série des] la probabilité qu’il y aura au moins une feule voix de plus en faveur de la vérité [(V^q)] est une série convergente [...] mais lorsque q est grand v&e restant les mêmes[.]” (p.4, p.6)

[Result 1]: The number of voters is $2q + 1$, and we are looking for the probability of the majority.
If $v > e$, then the probability that the decision will conform to the truth will increase steadily with increasing the number of voters;
The series of the probability that there will be at least one more voice in favor of the truth (V^q) is a convergent series [...] and hence when q is large v and e will remain the same, [and so the majority will be right].

“Cette première observation nous conduit d’abord à cette conséquence, que plus le nombre des Votans fera grand, plus il y a de probabilité que leur décision fera contraire à la vérité lorsque $e > v$, c’est-à-dire lorsqu’il y a probabilité que chacun en particulier se trompera; & si q est très-grand, cette probabilité pourra devenir très-grande, quoique la différence entre v&e soit très-petite.” (p.6)

[Result 2]: This observation leads us to the consequence, that the greater the number of voters, the greater the probability that their decision will be contrary to the truth if $e > v$, that is to say: There is a probability that everyone in particular will be deceived; If q is very large, this probability may become very great, although the difference between v and e might be very small.

Now, by help of our framing of the weak law of large numbers as a wise crowd effect, we can demonstrate the part on the convergence in Condorcets result 1 and 2: Assume $\{Z_1, Z_2, \ldots\}$ to be a set of independent and equally competent/incompetent voters regarding some fact z, where voting is considered as a binary process only, i.e. either $Z_i = z$ or $Z_i \neq z$ (for all $Z_i, Z_j \in \{Z_1, Z_2, \ldots\}$):

- $Pr(Z_i = z | Z_j = z) = Pr(Z_i = z)$
- $Pr(Z_i = z) = Pr(Z_j = z) = v$ (and $e = 1 - v$)

Then the weak law of large numbers states:

$$\text{For all } \varepsilon > 0: \lim_{q \to \infty} Pr\left(\left| \frac{Z_1 + \cdots + Z_q}{q} - v \right| < \varepsilon \right) = 1$$

Now, suppose $1 > v > e$, i.e. $v = 0.5 + \varepsilon_1$. Then $|(Z_1 + \cdots + Z_q)/q - v| < \varepsilon_1$ almost certainly if $q \to \infty$. And hence $(Z_1 + \cdots + Z_q)/q > 0.5$ almost certainly (with $q \to \infty$). Hence, also the majority, i.e. majority aggregation (oddness of q presupposed)

$$Z_{majority(q)} = z \text{ iff } \frac{|\{Z_i : 1 \leq i \leq q \text{ and } Z_i = z\}|}{|\{Z_i : \leq i \leq q \text{ and } Z_i \neq z\}|} > 0.5$$

is almost certainly right—we end up with the wise crowd effect:

$$\lim_{q \to \infty} Pr(Z_{majority(q)} = z) = 1 > v = Pr(Z_i = z)$$

The argument for a *madness* of the crowd effect (convergence part in result 2 of Condorcet) is analogous. If $0 < v < e$, i.e. $0 < v < 0.5$, then aggregating the votes of the incompetent jury members by help of a majority rule leads almost certainly to a false outcome if the number of jury members is arbitrary high. In this case the crowd almost certainly performs worse than the average individual.

We have seen in the citations of the *Essai* above, that Condorcet considers two relevant parameters in his model: v, the probability of an individual to judge correctly, and q, the number of individuals. In our framing also a third parameter is relevant for the investigations in the *Essai*, namely a parameter m for specifying the aggregation method. In the case discussed above, m is 0.5 for characterising *absolute* majority aggregation. However, also other values might be of interest as, e.g., $m = 2/3$ for a *qualified* majority, etc. As is pointed out by Daston (1988, p.351), from the result above "Condorcet concluded that the probability that the tribunal would render a correct decision in any given case could be increased by increasing either [q, m], or v. Much of the Essai is devoted to examining the consequences of allowing one of these to vary while the others remained constant". So, the

Essai engaged already systematically with what is called nowadays *applied epistemology* (see Goldman 2011a, p.32) or *institutional design* (see List 2011, p.221) in social epistemology.

The wise crowd effect stated in terms of the law of large numbers is a nice gauge for illustrating how the collective outperforms the individual in case that v, the individual competence, or m, the majority threshold, or g, the number of individuals is high enough. However, the model is only very coarse grained (for q arbitrary high, etc.). For this reason we also want to briefly discuss the recursive form of Condorcet's jury theorem, complementing the other parts of the results cited above.

We start with a simple comparison of absolute majority aggregation and the average individual: Assume a binary prediction task, and a group of predictors with the majority of a individuals (e.g. all predicting 1) and the minority of $i < a$ individuals (e.g. all predicting 0). The number of individuals in the collective is $g = a + i$. Now, let v be the probability that an individual predicts the correct value (all individuals have the same probability), and let the individuals' predictions be probabilistically independent. I.e.: The probability of 2 individuals predicting correctly equals v^2, $\ldots$, that of n individuals predicting correctly equals v^n. To use the same terms as Condorcet, let $e = 1 - v$ be the probability that an individual predicts incorrectly. Now, for any collective with size g and binary predictions, there are $2 \cdot \binom{g}{a}$ majorities with exactly a equivalent elements (e.g.: if $g = 3$, then the possible combinations of predictions are the following $2^g = 8$ sequences: $0,0,0$ and $0,0,1$ and $0,1,0$ and $0,1,1$ and $1,0,0$ and $1,0,1$ and $1,1,0$ and $1,1,1$; out of these exactly $2 \cdot \binom{2^g}{2^a}$ have a equal predictions with $a > g/2$; if $a = 3$, then it is $2 \cdot \binom{3}{3} = 2$, namely $0,0,0$ and $1,1,1$; if $a = 2$, then it is $2 \cdot \binom{3}{2} = 6$, namely the remaining sequences). So, the probability that a majority of a individuals of a collective with g individuals predicts correctly is:

$$2 \cdot \binom{g}{a} \cdot v^a \cdot e^i$$

The probability that such a majority predicts *in*correctly is:

$$2 \cdot \binom{g}{a} \cdot e^a \cdot v^i$$

Therefore, the probability for any particular majority (of one particular sequence) is:

$$\frac{2 \cdot \binom{g}{a} \cdot v^a \cdot e^i}{2 \cdot \binom{g}{a} \cdot v^a \cdot e^i + 2 \cdot \binom{g}{a} \cdot e^a \cdot v^i} = \frac{v^a \cdot e^i}{v^a \cdot e^i + e^a \cdot v^i}$$

We find the latter term also in (Condorcet 1785, p.11; see Daston 1988, p.350). In comparing it with the individual/average probability of a correct prediction v it holds that:

$$\frac{v^a \cdot e^i}{v^a \cdot e^i + e^a \cdot v^i} \begin{cases} > v, & \text{iff } 1 > v > 0.5 \\ = v, & \text{iff } v = 0.5 \\ < v, & \text{iff } 0 < v < 0.5 \end{cases}$$

If we want to compare the collective with a sub-collective, we can apply the same idea, but need to add more details. We do so again by combinatorial considerations: For illustrative purposes, let us assume that the true value is 1 and assume a collective with $g = 3$ individuals with the predictions Z_1, Z_2, Z_3. The probabilities of these predictions are depicted in table 12.1, the grey lines mark cases where a majority predicts correctly. A majority has at least $a = (g+1)/2$ members. In the case of $g = 3$ the cases are relevant where a majority with 2 members produces a correct prediction—there are $\binom{3}{2} = 3$ such cases—and a majority with 3 members produces such one—there is $\binom{3}{3} = 1$ such case. So, in case of $g = 3$ we can sum up the probabilities of a majority reaching a correct decision as $\binom{3}{2} \cdot v^2 e^1 + \binom{3}{3} \cdot v^3 e^0$ (see Estlund 1994, e.g.). Table 12.2 displays the case for a jury with $g = 5$. Here the probability for a majority reaching a correct decision is the sum of $\binom{5}{3} \cdot v^3 e^2$, $\binom{5}{4} \cdot v^4 e^1$, and $\binom{5}{5} \cdot v^5 e^0$. Now, as one can see, the general case for a Condorcet jury with the individual probability of making a correct prediction v ($e = 1 - v$) and g probabilistically independent members is:

Z_3	Z_2	Z_1		
0	0	0	$Pr: \; e \cdot e \cdot e \; =$	$v^0 e^3$
0	0	1	$Pr: \; e \cdot e \cdot v \; =$	$v^1 e^2$
0	1	0	$Pr: \; e \cdot v \cdot e \; =$	$v^1 e^2$
0	1	1	$Pr: \; e \cdot v \cdot v \; =$	$v^2 e^1$
1	0	0	$Pr: \; v \cdot e \cdot e \; =$	$v^1 e^2$
1	0	1	$Pr: \; v \cdot e \cdot v \; =$	$v^2 e^1$
1	1	0	$Pr: \; v \cdot v \cdot e \; =$	$v^2 e^1$
1	1	1	$Pr: \; v \cdot v \cdot v \; =$	$v^3 e^0$
				$\sum = 1$

Table 12.1: Probability distribution of a Condorcet jury with three members with individual probability of a correct (here assumed: 1) prediction v ($e = 1 - v$); the cases where a majority gets things right are marked grey.

$$v_g = \sum_{j=\frac{g+1}{2}}^{g} \binom{g}{j} \cdot v^j \cdot e^{g-j}$$

By replacing 'g' by '$g+2$' we get the probability of the next (odd numbered) Condorcet jury to make a correct prediction. Grofman, Owen, and Feld

Z_5	Z_4	Z_3	Z_2	Z_1	
0	0	0	0	0	$Pr : v^0e^5$
0	0	0	0	1	$Pr : v^1e^4$
0	0	0	1	0	$Pr : v^1e^4$
0	0	0	1	1	$Pr : v^2e^3$
0	0	1	0	0	$Pr : v^1e^4$
0	0	1	0	1	$Pr : v^2e^3$
0	0	1	1	0	$Pr : v^2e^3$
0	0	1	1	1	$Pr : v^3e^2$
0	1	0	0	0	$Pr : v^1e^4$
0	1	0	0	1	$Pr : v^2e^3$
0	1	0	1	0	$Pr : v^2e^3$
0	1	0	1	1	$Pr : v^3e^2$
0	1	1	0	0	$Pr : v^2e^3$
0	1	1	0	1	$Pr : v^3e^2$
0	1	1	1	0	$Pr : v^3e^2$
0	1	1	1	1	$Pr : v^4e^1$
1	0	0	0	0	$Pr : v^1e^4$
1	0	0	0	1	$Pr : v^2e^3$
1	0	0	1	0	$Pr : v^2e^3$
1	0	0	1	1	$Pr : v^3e^2$
1	0	1	0	0	$Pr : v^2e^3$
1	0	1	0	1	$Pr : v^3e^2$
1	0	1	1	0	$Pr : v^3e^2$
1	0	1	1	1	$Pr : v^4e^1$
1	1	0	0	0	$Pr : v^2e^3$
1	1	0	0	1	$Pr : v^3e^2$
1	1	0	1	0	$Pr : v^3e^2$
1	1	0	1	1	$Pr : v^4e^1$
1	1	1	0	0	$Pr : v^3e^2$
1	1	1	0	1	$Pr : v^4e^1$
1	1	1	1	0	$Pr : v^4e^1$
1	1	1	1	1	$Pr : v^5e^0$

$\longrightarrow$

$\sum = 1$

Table 12.2: Probability distribution of a Condorcet jury with five members with individual probability of a correct (here assumed: 1) prediction v ($e = 1 - v$); the cases where a majority gets things right are marked grey. The example is constructed in line with the remarks above based on (Estlund 1994)

(1983, p.256) transform these formulae to the following recursion formula: The probability of a jury with $g + 2$ probabilistically independent members is:

$$v_{g+2} = v_g + \binom{g}{\frac{g+1}{2}} \cdot \left(\underbrace{v^2 v^{\frac{g-1}{2}} e^{\frac{g+1}{2}}}_{t_1} - \underbrace{e^2 e^{\frac{g-1}{2}} v^{\frac{g+1}{2}}}_{t_2} \right)$$

Whether $v_{g+2} > v_g$ clearly depends on whether $t_1 > t_2$. Again, it holds:

$$v_{g+2} \begin{cases} > v_g, & \text{iff } 1 > v > 0.5 \\ = v_g, & \text{iff } v = 0.5 \\ < v_g, & \text{iff } 0 < v < 0.5 \end{cases}$$

By this we now have also a recursive form of a wise crowd effect, stating that in a Condorcet jury with equally competent and probabilistically independent members (competence $v > 0.5$), increasing the number of individuals g of the collective also increases the probability of the majority aggregating group. If we interpret the probability of a collective to predict correctly as its expectation value of the score $s_{majority}$, and the individual probability v as its expectation value of the score $s_{individual}$, then we can formulate the Condorcet wise crowd effect as:

$$\mathbb{E}[s_{majority}] > \mathbb{E}[s_{individual}]$$

And for comparison of collectives: If $collective_1$ and $collective_2$ are Condorcet juries with $collective_1 \subset collective_2$, then:

$$\mathbb{E}[s_{collective_2}] > \mathbb{E}[s_{collective_1}]$$

In the next section we are going to investigate conditions for wise crowd effects with non-probabilistic predictions.

12.3 Averaging Outperforms the Average

We now come to the case of non-probabilistic predictions. Here the task of the individuals and the collective is to predict an event outcome by help of providing a single value. To remain within our prediction setting as described in chapter 2, we assume that the event outcome is described by an element of $[0, 1]$ and that the prediction consists also of an element of $[0, 1]$. So, e.g., in the ox-weight-judgement competition discussed by Galton, the competition organisers might simply restrict judgements to an interval of 0 to 2000 pounds, and then normalise the single judgements via dividing the provided values by 2000. For the rest of this section we use the notation of chapter 2 and omit the index for the event, since we are considering only single events. So, Y is the event, $y \in [0, 1]$ is the true value representing the event outcome, $\mathcal{F} = \{f_1, \ldots, f_n\}$ is the set of individual predictions of Y, i.e. $f_i \in [0, 1]$. Furthermore, also in the loss function $\ell_i = \ell(f_i, y)$, and the score $s_i = 1 - \ell_i$ no reference is made to an event index.

It is interesting to note that also for non-probabilistic predictions a wise crowd effect can be characterised. Krogh and Vedelsby (1995) have found a very interesting way of describing the performance of a collective prediction in comparison to an *average* individual's prediction: Take the group's prediction of the value of an event Y_1 to be the average of the individuals' predictions (for the following see Krogh and Vedelsby 1995, pp.232f, we defined the average learner f_{av} for the dynamic setting with a series of events in definition 3.31 on p.105):

Definition 12.1 (Prediction of the Collective)**.**

$$f_{av} = \frac{\sum_{i=1}^{n} f_i}{n}$$

If we want to compare the collective's prediction (again, be aware that the collective prediction is the average of predictions, so there might be no actual individual making that prediction) with that of the individuals, then we cannot do this directly since the individuals' predictions may be heterogeneous. But we can define the notion of an *average individual* and compare the collective's prediction with that of an average individual via the errors in these predictions. For this purpose, we introduce a measure for the error of a prediction simply by help of the loss function ℓ. E.g., one might think of the squared difference of true event outcome and the predicted outcome as such a measure of error. First, we can take the error of an individual prediction f_i to be simply ℓ_i. Then we can define a measure for the *average* individuals' error just by calculating the average of the error of each individual (see Krogh and Vedelsby 1995, p.232)—note also here that there might be no individual in the setting which represents the *average* individual. We refer to the fictive *average* individual of a group $\mathcal{F} = \{f_1, \ldots, f_n\}$ with the help of '$f_{\varnothing\{1,\ldots,n\}}$':

Definition 12.2 (Prediction Error of Average Individual)**.**

$$\ell_{\varnothing\{1,\ldots,n\}} = \frac{\sum\limits_{i=1}^{n} \ell_i}{n}$$

And similar to the individuals' errors we measure the error of the collective's prediction simply via the loss function: ℓ_{av} (which is $\ell(f_{av}, y)$). So, we have:

- ℓ_i is the error of the individual prediction f_i.
- ℓ_{av} is the error of the collective prediction f_{av}.
- $\ell_{\varnothing\{1,\ldots,n\}}$ is the error of the (fictive) *average* individual of group $\mathcal{F} = \{f_1, \ldots, f_n\}$.

Now, given that the underlying loss function ℓ is convex, we get the following result:

Theorem 12.3 (The Crowd Beats the Average Law (see Page 2007, p.209; and Krogh and Vedelsby 1995, p.233))**.** *Given ℓ is convex, it holds:*

$$\ell_{av} \leq \ell_{\varnothing\{1,\ldots,n\}}$$

Proof. Let us assume that ℓ is convex. Just reformulate ℓ_{av} as defined in definition 12.1 as: $\ell(\frac{1}{n}f_1 + \cdots + \frac{1}{n}f_n, y)$. Likewise, reformulate $\ell_{\varnothing\{1,\ldots,n\}}$ as defined in definition 12.2 as: $\frac{1}{n}\ell(f_1, y) + \cdots + \frac{1}{n}\ell(f_n, y)$. Then it is easy to see that ℓ_{av} is the loss of a weighted average of predictions, whereas $\ell_{\varnothing\{1,\ldots,n\}}$ is the weighted average of the losses of these predictions. Hence, by definition 3.30 (convexity of ℓ), $\ell_{av} \leq \ell_{\varnothing\{1,\ldots,n\}}$. □

This theorem shows that the error of a prediction of the collective is equal to or smaller than the error of an *average* member of the collective, which is a very general positive feature of applying a meta method (namely averaging) in predicting the value of an event.

Now, recall our general characterisation of a wise crowd effect of section 12.1: Such an effect is at hand, if the collective prediction is better than an individual prediction or the prediction of another part of the collective. Clearly, what the *Crowd Beats the Average Law* aims at is to compare the collective prediction with that of the *average* individual. However, as stated up to now, we only know that aggregating predictions by averaging allows for being at least as accurate as the prediction of an average individual. We do not know whether averaging allows also for being *more accurate*, i.e. better, than the average individual. However, as the analysis of Krogh and Vedelsby (1995, p.232) shows, there is a measure for characterising cases where *averaging outperforms the average*, namely a measure for the *diversity* of the predictions of the individuals. The idea is to measure the degree of diversity of the prediction of an individual of a collective as dependent on its deviation from the average prediction. Note that ℓ_i measures the loss of prediction f_i compared to the *true value*. In contrast to this, the diversity measure we aim at measures the loss of a prediction compared to the *average predicted value*. So, if we take d_i to be a measure for the diversity of the prediction f_i, then d_i can be defined as: $d_i = \ell(f_i, f_{av})$ (recall that in contradistinction to this the individual error/loss ℓ_i was defined as $\ell(f_i, y)$). Now, what is relevant for characterising a wise crowd effect is not the measure for the diversity of an individual prediction, but a measure for the average diversity, i.e. the diversity of the (fictive) *average* individual. If we use '$d_{\varnothing\{1,\ldots,n\}}$' for this measure, we can define it as follows:

Definition 12.4 (Diversity of Average Individual)**.**

$$d_{\varnothing\{1,\ldots,n\}} = \frac{\sum_{i=1}^{n} \ell(f_i, f_{av})}{n}$$

Now, this measure allows for characterising a wise crowd effect. It holds:

Theorem 12.5 (The Diversity Prediction Theorem (see Page 2007, p.208; and Krogh and Vedelsby 1995, p.232))**.** *If ℓ is the quadratic error function ($\ell(x,y) = (x-y)^2$), then:*

$$\ell_{av} = \ell_{\varnothing\{1,\ldots,n\}} - d_{\varnothing\{1,\ldots,n\}}$$

Proof.

1. Assume that ℓ is the quadratic error function: $\ell(x,y) = (x-y)^2$
2. Then we get by definition 12.2: $\ell_{\varnothing\{1,\ldots,n\}} = \frac{\sum_{i=1}^{n}(f_i - y)^2}{n}$
3. Furthermore: $\ell_{av} = (f_{av} - y)^2$ where according to definition 12.1: $f_{av} = \frac{\sum_{i=1}^{n} f_i}{n}$
4. Also, by definition 12.4, we get: $d_{\varnothing\{1,\ldots,n\}} = \frac{\sum_{i=1}^{n}(f_i - f_{av})^2}{n}$

5. Now, let us resolve $d_{\varnothing\{1,\ldots,n\}}$ further:

$$\begin{aligned}
d_{\varnothing\{1,\ldots,n\}} &= \frac{\sum_{i=1}^{n} (\ f_i - f_{av}\ )^2}{n} \\
&= \frac{\sum_{i=1}^{n} (\ (f_i - y) - (f_{av} - y)\ )^2}{n} \\
&= \frac{\sum_{i=1}^{n} (f_i - y)^2 + (f_{av} - y)^2 - 2(f_{av} - y)(f_i - y)}{n} \\
&= \underbrace{\frac{\sum_{i=1}^{n} (f_i - y)^2}{n}}_{=\ell_{\varnothing\{1,\ldots,n\}}} + \underbrace{\underbrace{\frac{\sum_{i=1}^{n} (f_{av} - y)^2}{n}}_{=n\cdot(f_{av}-y)^2/n}}_{=\ell_{av}} - \\
&\quad - \underbrace{\frac{\sum_{i=1}^{n} 2(f_{av} - y)(f_i - y)}{n}} \\
&\quad \underbrace{=2(f_{av} - y)\ \cdot\ \underbrace{\sum_{i=1}^{n} \frac{f_i - y}{n}}_{=\underbrace{\left(\sum_{i=1}^{n} f_i\right)/n}_{=f_{av}} - \underbrace{n\cdot y/n}_{=y}}}_{=2(f_{av}-y)^2=2\ell_{av}} \\
&= \ell_{\varnothing\{1,\ldots,n\}} + \ell_{av} - 2\ell_{av} \quad = \quad \ell_{\varnothing\{1,\ldots,n\}} - \ell_{av}
\end{aligned}$$

6. Hence:

$$\ell_{av} \quad = \quad \ell_{\varnothing\{1,\ldots,n\}} \ - \ d_{\varnothing\{1,\ldots,n\}}$$

□

Theorem 12.5 states that generally, the lower the average error or the higher the diversity within a collective, the lower the error of the collective's prediction. From this the following characterisation of a wise crowd effect follows immediately:

Theorem 12.6 (Averaging Wise Crowd Effect)**.** *Under the conditions that*

- *ℓ is the quadratic error function, and*

- *that the collective is diverse in the sense that* $d_{\varnothing\{1,\dots,n\}} > 0$

aggregating the individual predictions by averaging, i.e. f_{av}*, is guaranteed to provide a better prediction than that of the (fictive)* average *individual:*

$$\ell_{av} < \ell_{\varnothing\{1,\dots,n\}}$$

The condition regarding $d_{\varnothing\{1,\dots,n\}}$ is quite weak. Note that given the quadratic error loss function, according to definition 12.4 this condition is not satisfied, i.e. $d_{\varnothing\{1,\dots,n\}} = 0$, only if all individuals make the exact same predictions. In all other cases d is strictly positive, and hence a wise crowd effect is guaranteed.

What about the condition that ℓ is the quadratic error function? This condition relevantly entered the proof of theorem 12.5. It allowed us to define a measure for diversity which provided a characterisation of a wise crowd effect. However, we can also generalise this assumption: A wise crowd effect is guaranteed, for any *strictly convex* loss function ℓ, as long as not all individuals of the collective make the same prediction:

Theorem 12.7 (The Crowd *Really* Beats the Average Law)**.** *Given* ℓ *is strictly convex and* $d_{\varnothing\{1,\dots,n\}} > 0$ *(i.e. for some* $f_i, f_j \in \mathcal{F}$ *it holds that* $f_i \neq f_j$*. Then:*

$$\ell_{av} < \ell_{\varnothing\{1,\dots,n\}}$$

Proof. The proof is analogous to that of theorem 12.7. One only needs to substitute 'strictly convex' for 'convex' and '$<$' for '$\leq$'; strict convexity is analogously defined to convexity as in definition 3.30, just '$\leq$' is to be replaced by '$<$'. The assumption that at least two individuals of the collective make a different prediction guarantees that the average prediction deviates from the prediction of at least one individual. □

The quadratic error function is strictly convex ($(w \cdot x_1 + (1-w) \cdot x_2 - y) < w \cdot (x_1 - y)^2 + (1-w) \cdot (x_2 - y)^2$ for $0 < w < 1$), and hence proven to allow for a wise crowd effect already according to theorem 12.7. Also a normalised exponential error function which takes the absolute difference between f_i and y in the exponent is strictly convex, and hence allows for a wise crowd effect. But what about not strictly convex loss functions? So, e.g., what about the absolute loss $\ell(x,y) = |x-y|$, or what about concave loss functions as, e.g., a logarithmic error function $\ell(x,y) = \log_{\varepsilon}(max(\varepsilon, |x-y|))$ (with any $\varepsilon \in [0,1]$)?

Let us begin with the absolute loss: It is not hard to find an example where there is some diversity within the collective, i.e. $d_{\varnothing\{1,\dots,n\}} > 0$, and still $\ell_{av} = \ell_{\varnothing\{1,\dots,n\}}$. Consider the case where $f_1 = 0.25, f_2 = 0.15, y = 0.35$. Then, given the absolute loss, $d_{\varnothing\{1,\dots,n\}} = 0.05$, and $\ell_{\varnothing\{1,\dots,n\}} = (0.10 + 0.20)/2 = 0.15$; furthermore, $f_{av} = (0.25 + 0.15)/2 = 0.20$. Hence, $\ell_{av} = |f_{av} - y| = 0.15$ which equals $\ell_{\varnothing\{1,\dots,n\}}$. Hence, given the absolute loss function, not always the collective performs better than the *average* individual. So, $d_{\varnothing\{1,\dots,n\}} > 0$ does not characterise a wise crowd effect. However, as Lyon (forthcoming) discusses, there is another condition which allows for such a characterisation: Considering our example one can notice that both individual predictions were below the true value. Now, strict positive diversity suffices only for better performance of the collective in the light of the absolute loss function, if there is diversity in the individual predictions in the sense that not all of them are either below or above the true value. I.e., if some individual predictions are *over*estimating the event outcome, and some are *under*estimating it, then also the absolute loss function allows for a wise crowd effect (see Lyon forthcoming, sect.2). This condition for a positive effect of averaging is well-known in the psychological literature and is called '*bracketing*': "As the actual rate of bracketing [the truth] increases, so does the power of averaging" (see Larrick and Soll 2006, p.112). In line with Lyon (forthcoming), we do not aim at a quantitative description of this effect here, but only at a qualitative one:

Theorem 12.8 (Another Averaging Wise Crowd Effect)**.** *Under the conditions that*

- *ℓ is convex (e.g., the absolute loss $\ell(x, y) = |x - y|$), and*
- *that the collective is diverse in the sense that some individuals f_i and f_j bracket the truth y, i.e. there are $f_i, f_j \in \mathcal{F}$ such that $f_i < y < f_j$*

aggregating the individual predictions by averaging, i.e. f_{av}, is guaranteed to provide a better prediction than that of the (fictive) average *individual:*

$$\ell_{av} < \ell_{\varnothing\{1,\dots,n\}}$$

Proof. For the case of a strictly convex loss function ℓ, it suffices to note that $f_i < y < f_j$ implies that $d_{\varnothing\{1,\dots,n\}} > 0$, and hence theorem 12.6 applies.

For the case of the absolute loss function ℓ, we construct the average in three steps: First, we average among all overestimations. Second, we average among all underestimations including those predictions which equal the truth. Since both sets are disjoint, we can finally construct the average by weighted averaging these averages: Let $\mathcal{F}_o$ be the set of overestimations, i.e.: $\mathcal{F}_o = \{f_i \in \mathcal{F} : f_i > y\}$. Let $\mathcal{F}_u$ be the set of underestimations and perfect predictions, i.e.: $\mathcal{F}_u = \{f_i \in \mathcal{F} : f_i \leq y\}$. Now define f_{av_o} as the average of $\mathcal{F}_o$: $f_{av_o} = \frac{\sum_{f_i \in \mathcal{F}_o} f_i}{|\mathcal{F}_o|}$. Similarly, $f_{av_u} = \frac{\sum_{f_i \in \mathcal{F}_u} f_i}{|\mathcal{F}_u|}$. Since for all $f_i \in \mathcal{F}_o$:

$f_i > y$, it holds: $\ell_{av_o} = \frac{|\sum_{f_i \in \mathcal{F}_o}(f_i) - |\mathcal{F}_o| \cdot y|}{|\mathcal{F}_o|} = \frac{\sum_{f_i \in \mathcal{F}_o}(|f_i - |\mathcal{F}_o| \cdot y|)}{|\mathcal{F}_o|} = \ell_{\varnothing \mathcal{F}_o}$. Hence $\ell_{av_o} = \ell_{\varnothing \mathcal{F}_o}$. Analogously it holds for f_{av_u}: $\ell_{av_u} = \ell_{\varnothing \mathcal{F}_u}$. Since $f_i < y < f_j$ for some $f_i, f_j \in \mathcal{F}$, it holds that $f_{av_u} < y < f_{av_o}$. Now we weight the average losses to get the overall losses. The overall loss of the *average* individual amounts to: $\ell_{\varnothing \mathcal{F}_o \cup \mathcal{F}_u} = \ell_{\varnothing \mathcal{F}} = \ell_{\varnothing\{1,\ldots,n\}} = \frac{|\mathcal{F}_o|}{|\mathcal{F}|} \cdot \ell_{\varnothing \mathcal{F}_o} + \frac{|\mathcal{F}_u|}{|\mathcal{F}|} \cdot \ell_{\varnothing \mathcal{F}_u} = \frac{|\mathcal{F}_o|}{|\mathcal{F}|} \cdot \ell_{av_o} + \frac{|\mathcal{F}_u|}{|\mathcal{F}|} \cdot \ell_{av_u}$. However, the overall loss of the collective prediction amounts to: $\ell_{av} = \left| \frac{|\mathcal{F}_o|}{|\mathcal{F}|} \cdot \ell_{av_o} + \frac{|\mathcal{F}_u|}{|\mathcal{F}|} \cdot \ell_{av_u} - y \right|$. Hence, $\ell_{av} < \ell_{\varnothing\{1,\ldots,n\}}$. □

Diversity in the sense of bracketing ($f_i < y < f_j$ for some f_i, f_j of the collective) is a stronger condition than diversity in the sense of a deviation of the average ($d_{\varnothing\{1,\ldots,n\}} > 0$). However, also for the absolute loss function it generally holds that the *average* individual does not outperform the collective. This follows immediately from theorem 12.3, since this loss is also convex, though clearly not strictly convex. It can be also seen by help of the subadditivity property of the absolute value function, according to which $|x + y| \leq |x| + |y|$: We know that $\ell_{av} = |\frac{1}{n} \cdot \sum_{i=1}^{n}(f_i) - y| = \frac{1}{n} \cdot |\sum_{i=1}^{n}(f_i - y)|$. Furthermore, $\ell_{\varnothing\{1,\ldots,n\}} = \frac{1}{n} \cdot \sum_{i=1}^{n}(|f_i - y|)$, and hence by subadditivity: $\ell_{av} \leq \ell_{\varnothing\{1,\ldots,n\}}$.

This is different with strictly convex loss functions: It is easy to see that these do not only *not* guarantee a wise crowd effect, but on the contrary, once $d_{\varnothing\{1,\ldots,n\}} > 0$, they even guarantee an *anti*-wise crowd effect (see Lyon forthcoming, sect.2). This follows analogously to before:

Theorem 12.9 (The Average Beats the Crowd Law)**.** *Given ℓ is strictly concave and $d_{\varnothing\{1,\ldots,n\}} > 0$ (i.e. for some $f_i, f_j \in \mathcal{F}$ it holds that $f_i \neq f_j$. Then:*

$$\ell_{av} > \ell_{\varnothing\{1,\ldots,n\}}$$

Proof. The proof is analogous to that of theorem 12.7. One only needs to substitute 'strictly concave' for 'convex' and '$>$' for '$\leq$'; strict concavity is analogously defined to convexity as in definition 3.30, just '$\leq$' is to be replaced by '$>$'. Again, the assumption that at least two individuals of the collective make a different prediction guarantees that the average prediction deviates from the prediction of at least one individual. □

The results of this section are summed up in table 12.3.

Note that we can also express the performance of the collective via the score s. In case of a wise crowd effect, it holds:

$$s_{av} > s_{\varnothing\{1,\ldots,n\}}$$

This follows from the theorems above, since $s = 1 - \ell$. Furthermore, note that up to now we have characterised wise crowd effects with respect to

loss function ℓ	**condition**	**aggregation method**	**note**
strictly convex	diversity in the sense of a deviation of the average: $d_{\varnothing\{1,\ldots,n\}} > 0$	averaging	
absolute	diversity in the sense of bracketing: $f_i < y < f_j$ for some f_i, f_j of the collective	averaging	
strictly concave	no diversity in the sense of a deviation of the average at all: $d_{\varnothing\{1,\ldots,n\}} = 0$	averaging	in order to avoid an *anti*-wise crowd effect

Table 12.3: Summary of conditions and aggregation methods for wise crowd effects or avoidance of *anti*-wise crowd effects in non-probabilistic predictions.

the *average* individual only. But what about other individuals? In particular one might wonder what the conditions for a wise crowd effect with respect to the *best* individuals of the collective are. Now, in the case of single predictions it seems that there is little to say about this. A trivial condition which allows for such a wise crowd effect is, e.g., the condition that all individuals are equally accurate or inaccurate. However, this is quite a restriction. In the next section we outline how the theory of meta-induction allows for less restrictive conditions in order to guarantee wise crowd effects with respect to best individuals of a collective.

12.4 Meta-Inductive Crowdsourcing

Recall, a wise crowd effect is defined by better performance of the collective than a component of it. In section 12.1 we have seen that there is empirical evidence that an individual as crowd within performs better than the individual as an individual, if it is supposed to make several estimations. In section 12.2 we have seen that a crowd performs estimatedly better than a subcrowd or the average (or the best) individual of the crowd in a binary (classificatory) prediction task, if the predictions of the individuals of the crowd are diverse in the sense of probabilistically independent, and if the individuals are equally minimally competent, i.e. better predictors than a fair coin. In section 12.3 we have seen that a crowd performs *actually* better than the average individual of the crowd in a regression prediction task, if at least two predictions of the individuals of the crowd are different, and a *strictly* convex loss function is used. This result can be expanded to using the absolute loss function, if some individual predictions bracket the true value.

Now, these wise crowd effects are mainly about comparing the crowd with the *average* individual. In case of Condorcet juries, all individuals are equally competent in predicting, hence the *average* individual is also the *best* individual. The condition for equal competence can be relaxed, but still, then the crowd is only expected to beat the best individual, but it is not guaranteed to do so. Here we want to interpret meta-induction such that it characterises a "long run" wise crowd effect ensuring *actual* better performance compared to the *best* individuals of the group.

Here is a quite simple implementation: Let us consider the attractivity weighting meta-inductive learning algorithm f_{ami} as defined in definition 3.39. This learner keeps track of the success rates of each individual as well as its own; only those individuals are attractive to it, which have a higher success rate. It makes a prediction by arithmetically weighting the individual attractive predictions, where the weights are normalised attractivities (difference between the past success rate of the individual and the past success rate of f_{ami}). We have seen that f_{ami} is long run access optimal, since it is a no-regret algorithm (theorem 3.40). In case no individual is attractive, it uses a fallback strategy. The fallback strategy in definition 3.39 is averaging among the whole group. However, also other fallback strategies allow for optimality as, e.g., averaging among up to now two best individuals' predictions. In what follows, we assume that f_{ami} uses this fallback strategy, and provide also a further specification of which of the best individuals should be used for averaging in the fallback case.

According to theorem 3.40, the regret (difference of the cumulative loss) of f_{ami} to not have chosen the prediction f_i is: $aregret_{\langle ami,i\rangle,t} \leq \sqrt{n \cdot t}$. This means for the long run success rates (score averaged according to the number of rounds, i.e. the cumulative success divided by the number of rounds): $\lim_{t\to\infty}(succ_{ami,t} - succ_{i,t}) \geq 0$. So, f_{ami} cannot be outperformed in the long run by any individual of the group, neither the average individual nor the best one (if there is such a best one). However, for a wise crowd effect we need more: We need not only that f_{ami} cannot be outperformed, but outperforms itself the average or best individual. Outperforming the average individual is guaranteed, as long as in the long run the success rate of *the average* individual differs from that of the *best* one: In this case theorem 3.40 shows that f_{ami}'s success rate approaches that of the best individual and hence f_{ami} de facto outperforms the average individual. But how about our aim of showing that meta-inductive aggregation in the long run also outperforms the best individual(s)? Clearly, this holds not generally. So, e.g., in case there is only one best individual in the setting, then f_{ami} cannot outperform the best individual, since in the long run it will just imitate the best individual, i.e. copy its prediction. However, putting forward a strong diversity constraint allows for a wise crowd effect: If we assume that all prediction methods are always either over- or underestimating (including the case of a perfect estimation), and if we assume that for any overestimator there is an equal underestimator in the setting, then f_{ami} can be shown to even "beat" the experts. The diversity condition is quite strong and clearly not very realistic. However, we aim here only at a general conclusion, namely that diversity can be cashed out for a collective to even perform better than the best individual of the collective. The diversity assumption might be interpreted as *method-bracketing* or *systematic bracketing*. Here is, how we can use theorem 3.40 to proof that diversity in the sense of systematic bracketing allows for "beating" the experts:

Theorem 12.10 (Meta-Induction Beats the Expert)**.** *Let G be a prediction game with truth $\mathcal{Y}$ and the set of prediction methods $\mathcal{F}$. Furthermore, let $\mathcal{F}$ be diverse in the sense that there are several "best" experts allowing for* systematic bracketing*:*

- *For a best $1 \leq i \leq n$ there is a best $1 \leq j \leq n$ such that for all t:*

$$f_{i,t} - y_t = y_t - f_{j,t}$$

Furthermore, let $succ$ be defined on basis of a convex loss function ℓ. Then for all $1 \leq i \leq n$:

$$\lim_{t\to\infty}(succ_{ami,t} - succ_{i,t}) > 0$$

… as long as there are no long-run perfect predictors in $\mathcal{F}$ (i.e. there is no $f_i \in \mathcal{F}$ such that $\lim_{t\to\infty} succ_{i,t} = 1$)

Proof. By theorem 3.40 we know that $aregret_{\langle ami,i\rangle,t} \leq \sqrt{n \cdot t}$ for any $1 \leq i \leq n$. This means that $\sum_{u=1}^{t} \ell_{ami,t} \leq \sum_{u=1}^{t} \ell_{i,t} + \sqrt{n \cdot t}$. Hence:

$$\sum_{u=1}^{t} s_{ami,t} \geq \sum_{u=1}^{t} s_{i,t} - \sqrt{n \cdot t}$$

Now, by systematic bracketing we know that for every overestimation there is an equal underestimation. Let us denote a, up to t, best predictor which is an overestimator with 'b_o', and the respective best underestimator with 'b_u'. Now, as long as b_o and b_u are not perfect predictors, averaging among b_o and b_u will be closer to the true value y_t than any of them is. Hence, at round $t+1$, $s_{ami,t+1} > s_{b_o,t+1}$ and $s_{ami,t+1} > s_{b_u,t+1}$. Now, let us assume that $\varepsilon > 0$ is the surplus of the score that f_{ami} gains at $t+1$. We can choose ε to be the smallest surplus score gained by f_{ami} at any round, ignoring rounds with perfect predictions of the best predictor. The assumption that there is no long-run perfect predictor implies that—also for the long run—there always will be such b_os and b_us allowing for a ε surplus. So, generally we can assume that up to round t, f_{ami} gains $t \cdot \varepsilon$ compared to the respective b_o and b_u. Hence:

$$\sum_{u=1}^{t} s_{ami,t} \geq \sum_{u=1}^{t} s_{b_{o/u},t} - \sqrt{n \cdot t} + t \cdot \varepsilon$$

Now, if we average over this score, i.e. divide by t, we get:

$$succ_{ami,t} \geq succ_{b_{o/u},t} - \sqrt{\frac{n}{t}} + \varepsilon$$

And hence for any $1 \leq i \leq n$:

$$\lim_{t \to \infty} (succ_{ami,t} - succ_{i,t}) > 0$$

□

Theorem 12.10 shows that meta-induction guarantees a long run wise crowd effect even if we compare the collective with the best individuals of the collective. We must admit that the diversity assumption used for proving this result is very strong. However, we think that the result is still interesting for the following reasons: First of all, we think that such a wise crowd effect can be also guaranteed with weaker assumptions as, e.g., assuming that the average success of overestimators equals that of underestimators. Second, also the diversity assumption of probabilistic independence in Condorcet juries is quite strong. And third, meta-induction really adds a new element in the following sense: Clearly, if we apply ordinary averaging as described in the preceding section to the best predictions and assume, as our *systematic bracketing* assumption implies, that the best predictions are bracketing the true value, then also ordinary averaging outperforms the predictions of the best individuals. But note, neither meta-induction nor ordinary averaging is supposed "to know" which predictions in a round will be the best. Rather, meta-induction as well as ordinary averaging have to be applied on the set of all individuals of the collective. And there, ordinary averaging is not guaranteed to outperform the best individuals, even if their predictions equally bracket the true value. Consider, e.g., a collective with the individual predictions $0.1, 0.45, 0.55, 0.6$ and assume that the true value $y = 0.5$. Then the average of the predictions is 0.425 and hence outperformed by the best predictions 0.45 and 0.55. This, although the best predictions equally bracket the true value. In contrast to this, as theorem 12.10 shows, taking on a dynamic stance allows for proving that aggregating individual predictions in a success-based way even outperforms the best predictors, in case the setting is diverse enough.

We can summarise the main results on wise crowd effects of this chapter as in table 12.4: In the static non-probabilistic setting a wise crowd effect with regards to the average individual is guaranteed given a convex loss function and bracketing in the sense that at least some individuals of the collective over-, and some of them underestimate the true value. In the static probabilistic setting a wise crowd effect with respect to any subcollective, the average, and the best individuals is *expected*, if the individuals of the collective are equally competent and probabilistically independent. Finally, in the dynamic non-probabilistic setting a wise crowd effect with respect to the best individuals is guaranteed for the long run, given a convex loss function and systematic bracketing.

setting	**condition(s)**	**aggregation method**	**wise crowd effect**
static non-probabilistic	diversity in form of bracketing, convex loss (strictly convex or absolute loss)	averaging	$s_{crowd} > s_{\varnothing\{1,\ldots,n\}}$
static probabilistic	competence and diversity in form of prob. indep.	majority	$\mathbb{E}[s_{crowd}] > \mathbb{E}[s_{best}]$
dynamic non-probabilistic	diversity in form of systematic bracketing, convex loss	meta-inductive weighting	$\lim_{t\to\infty}\left(\frac{\sum_{u=1}^{t} s_{crowd,u}-s_{best,u}}{t}\right) > 0$

Table 12.4: Overview of different wise crowd effects in different settings: s_{crowd} is the score of the collective prediction, s_{best} is the score of the best prediction, $s_{\varnothing\{1,\ldots,n\}}$ is the score of the (fictive) prediction of the (fictive) average individual.

Conclusion

In this chapter a short summary of the results of this book is given in form of an overview of the general argumentation hierarchy.

If one had to summarise the line of reasoning of this book in a brief statement, it would be: If we allow for optimality as an epistemic end ($\mathcal{O}$) not only in the practical realm, but also in the theoretical one, then we can engineer a solution to the problem of epistemic justification, namely that of relative justification (J_r) via meta-induction. Schematically:

$$\mathcal{O}(optimality) \;\Rightarrow\; J_r(induction)$$

Still very roughly summarising, but a little bit more explicit, we take the following argumentation hierarchy as *the* upshot of our investigation.

Main argument of chapter 1: An absolute notion of justification leads to inductive scepticism.

Chpt.1-1 The absolute notion of epistemic justification constrained as being non-sceptic, non-foundational, non-coherentist, and non-infinitist is inconsistent—this is the problem of epistemic justification.
$\neg(S)\&\neg(F)\&\neg(C)\&\neg(I) \;\Rightarrow\; \lightning$

Chpt.1-2 Hence, at least one of scepticism or foundationalism or coherentism or infinitism is true. (from Chpt.1-1)
$(S) \vee (F) \vee (C) \vee (I)$

Chpt.1-3 Classical foundationalism, coherentism, and infinitism put forward epistemic ends which characterise an absolute notion of justification: truth, truth preservation, probability preservation/increase.
$(F) \vee (C) \vee (I) \;\Rightarrow\; \mathcal{O}(\text{truth})$

Chpt.1-4 However, it is impossible to establish the truth, truth preservation, or probability preservation/increase of important epistemic

principles of justification as, e.g., the principle of induction; such principles or inferences are no necessary and adequate means for these ends. (Chpt.2-1)
$\neg\Box(induction \rightarrow truth)$

Chpt.1-5 Hence, scepticism regarding an absolute justification of such epistemic principles follows. (from Chpt.1-2–Chpt.1-4 and justification via epistemic means-ends)
(S) regarding *absolute* justification of *induction*, i.e.:
$\neg J_a(induction)$

Main argument of chapter 2: In machine learning there is a modern analogue to classical inductive scepticism as described by the case of a *Cartesian daemon*, namely spam.

Chpt.2-1 That important epistemic principles of justification are no necessary means for the classical epistemic ends follows from classical scepticism: The case of a Cartesian daemon illustrates the possibility of important epistemic principles like induction to fail regarding the classical epistemic ends.
$daemon \;\Rightarrow\; \Diamond(induction \,\&\, \neg truth)$

Chpt.2-2 The classical epistemic ends truth, etc., are about facts, facts are described as events, so the epistemic task is to make inferences or predictions about events; this can be represented by the framework of prediction games.

Chpt.2-3 One branch of machine learning is particularly concerned with a structural analogue to the Cartesian daemon, namely adversarial supervised passive online learning as is used in spam detection.
$spam \;\Leftrightarrow\; daemon$ (by help of Chpt.2-1–Chpt.2-2)

Main argument of chapters 3 to 4: The modern analogue to inductive scepticism proves to still allow for guaranteed optimality via meta-induction.

Chpt.3-1 Spam detection or adversarial supervised passive online learning comes in two forms: as classification (discrete) task and as regression (continuous) task.
$spam \;\Leftrightarrow\; (spam_{class} \vee spam_{regr})$

Chpt.3-2 In case of online regression there is an inference or prediction method which is long run optimal relative to all available inference or prediction methods.
$spam_{regr} \;\Rightarrow\; \Box(meta\text{-}induction \rightarrow optimality)$

Chpt.4-3 In case of online classification there is no inference or prediction method which is strictly speaking optimal in the above sense, but

one which is expected to be optimal in this sense.
$spam_{class} \Rightarrow \Box(meta\text{-}induction \rightarrow optimality)$

Chpt.4-4 Hence, even in the very sceptic scenario induction proves to allow for optimal inferences or predicitons.
(from Chpt.3-1–Chpt.4-3)
$spam \Rightarrow \Box(meta\text{-}induction \rightarrow optimality)$

Main argument of chapter 5: Putting forward an optimality constraint for a relative notion of justification allows for an a priori justification of meta-induction, and an a posteriori justification of induction.

Chpt.5-1 Putting things together shows that even in case of a Cartesian daemon meta-inductive inference is optimal.
(from Chpt.2-3 and Chpt.4-4)
$daemon \Rightarrow \Box(meta\text{-}induction \rightarrow optimality)$

Chpt.5-2 Epistemic engineering and meta-induction puts forward optimality as an epistemic end—this is the relevant end for the relative notion of justification.
$\mathcal{O}(optimality)$

Chpt.5-3 Hence, we need not be sceptic regarding a relative notion of justification. (from Chpt.5-1 and Chpt.5-2)
$\neg(S)$ regarding *relative* justification, i.e.:
$J_r(meta\text{-}induction)$

Chpt.5-4 Now, in past inductive methods fared best, hence up to now meta-induction chooses induction as adequate means.
$meta\text{-}induction \Leftrightarrow induction$

Chpt.5-5 Hence, up to now we also need not be sceptic regarding induction. (from Chpt.5-1–Chpt.5-4)
$\neg(S)$ regarding *relative* justification of *induction*, i.e.:
$J_r(induction)$

Main argument of chapter 6: The relative justification of induction cannot be transferred to a relative justification of anti-induction in a Goodmanian manner since the underlying principle of language transformation is self-defeating.

Chpt.6-1 According to Goodman's new riddle of induction, language transformation allow for transferring justification of inductive inferences to a justification of anti-inductive inferences.
$J(induction)$ & $J(lang\text{-}trans) \Rightarrow J(anti\text{-}induction)$

Chpt.6-2 So, once language transformations are justified, we inherit relative justification of anti-induction. (from Chpt.6-1, and Chpt.5-5)
$J(lang\text{-}trans) \Rightarrow J_r(anti\text{-}induction)$

Chpt.6-3 However, allowing for language transformations is self-defeating.
$J(lang\text{-}trans) \Rightarrow \neg J(lang\text{-}trans)$

Chpt.6-4 Hence, the relative justification of induction cannot be transferred to anti-induction by help of language transformations.
$\not\Rightarrow J_r(anti\text{-}induction)$ (from Chpt.6-1–Chpt.6-3)

Main argument of chapters 7 to 8: So, whereas deduction allows for absolute justification, induction allows at least for relative justification. Putting forward simplicity constraints for optimisation, also abductive reasoning can be justified.

Chpt.8-1 Although there are some caveats, already by definition deductive principles and inferences are absolutely justified.
$J_a(deduction)$

Chpt.8-2 From the arguments above we know that the principle of induction is relatively justified. (from Chpt.5-5)
$J_r(induction)$

Chpt.7-3 By putting forward as epistemic end not only optimality regarding truth, but also optimality regarding truth and simplicity, the relative justification of induction can be transferred also to such a justification of abduction.
$J_{r'}(abduction)$

Chpt.7-4 Hence, optimisation or meta-inductive epistemic engineering allows for non-scepticism regarding the three common types of inferences: deduction, induction, and abduction.
(from Chpt.8-1–Chpt.7-3)
$\neg(S)$ regarding $deduction, induction, abduction$

Main argument of chapters 9 to 12: Meta-induction can be employed not only for justifying classical sources of knowledge, but also social ones: testimony, peer disagreement, and judgement aggregation.

Chpt.9-1 Meta-induction allows for relative justification of Hume's reliabilism regarding testimony;
J_r(THume)

Chpt.10-2 also for relative justification of the equal weight view regarding epistemic peer disagreement;
J_r(EWV)

Chpt.11-3 and also for relative justification of success-based scoring in probabilistic judgement aggregation—so-called arithmetic meta-inductive probability aggregation.
$J_r(\mathrm{AMI}^p)$

Chpt.11-4 Hence, optimisation or meta-inductive epistemic engineering allows for non-scepticism regarding the three social sources of knowledge: testimony, disagreement, and judgement aggregation. (from Chpt.9-1–Chpt.11-3)
$\neg(S)$ regarding $testimony, disagreement, judgement\ aggregation$

Figure 12.2 depicts the rough argumentation hierarchy of this investigation.

It seems that being "satisfied with the best" allows for justifying a wide range of classical and social sources of knowledge.

"*Sometimes the simplest tastes* are *the best.*"

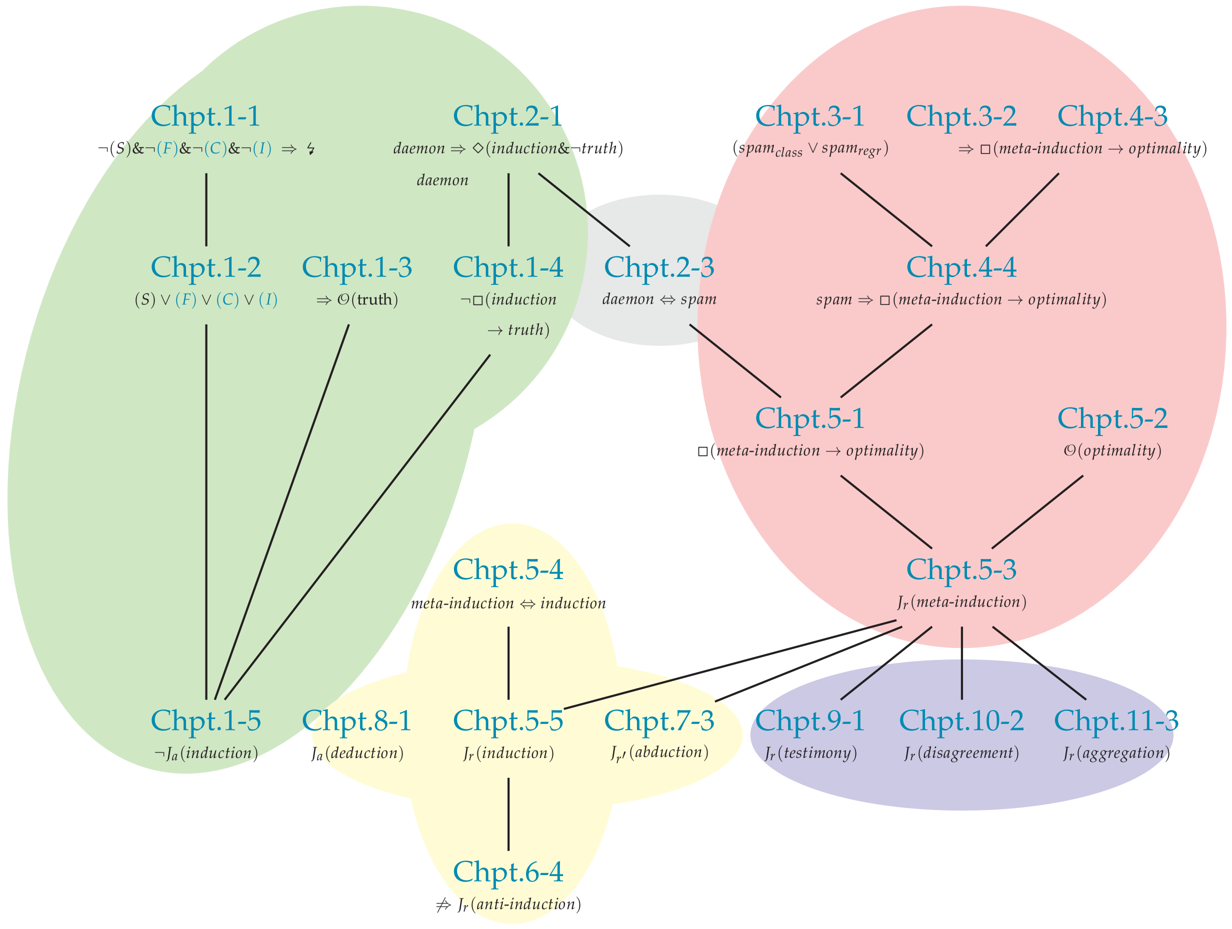
Chpt.1-1
$\neg(S)\&\neg(F)\&\neg(C)\&\neg(I) \Rightarrow ϟ$
Chpt.2-1
$daemon \Rightarrow \Diamond(induction\&\neg truth)$
$daemon$
Chpt.3-1
$(spam_{class} \vee spam_{regr})$
Chpt.3-2
Chpt.4-3
$\Rightarrow \Box(meta\text{-}induction \rightarrow optimality)$
Chpt.1-2
$(S) \vee (F) \vee (C) \vee (I)$
Chpt.1-3
$\Rightarrow \mho(\text{truth})$
Chpt.1-4
$\neg\Box(induction$
$\rightarrow truth)$
Chpt.2-3
$daemon \Leftrightarrow spam$
Chpt.4-4
$spam \Rightarrow \Box(meta\text{-}induction \rightarrow optimality)$
Chpt.5-1
$\Box(meta\text{-}induction \rightarrow optimality)$
Chpt.5-2
$\mho(optimality)$
Chpt.5-4
$meta\text{-}induction \Leftrightarrow induction$
Chpt.5-3
$J_r(meta\text{-}induction)$
Chpt.1-5
$\neg J_a(induction)$
Chpt.8-1
$J_a(deduction)$
Chpt.5-5
$J_r(induction)$
Chpt.7-3
$J_{r'}(abduction)$
Chpt.9-1
$J_r(testimony)$
Chpt.10-2
$J_r(disagreement)$
Chpt.11-3
$J_r(aggregation)$
Chpt.6-4
$\not\Rightarrow J_r(anti\text{-}induction)$

Figure 12.2: Rough argumentation hierarchy of our investigation; green: the problem of epistemic justification and induction; red: meta-induction; grey: mapping of the classical problem to meta-induction (mainly in part I); yellow: application of meta-induction to the classical epistemic realm (part II); blue: application of meta-induction the social epistemic realm (part III);

References

Adler, Jonathan E. (2012). "Epistemological Problems of Testimony". In: *The Stanford Encyclopedia of Philosophy (Winter 2010 Edition)*. Ed. by Zalta, Edward N.

Aikin, Scott F. (2011). *Epistemology and the Regress Problem*. New York: Routledge. DOI: 10.4324/9780203833247.

Aristotle (1957). *Aristotle's Prior and Posterior Analytics*. Ed. by Ross, William David. Oxford: Oxford University Press.

— ed. (1993). *Metaphysics. Books Gamma, Delta, and Epsilon*. Oxford: Clarendon Press.

Arnold, Eckhart (2010). "Can the Best-Alternative Justification Solve Hume's Problem? On the Limits of a Promising Approach". In: *Philosophy of Science* 77.4, pp. 584–593. DOI: 10.1086/656010.

Arrow, Kenneth Joseph (1963). *Social Choice and Individual Values*. 2nd Edition. Yale: Yale University Press.

Atkinson, David and Peijnenburg, Jeanne (2009). "Justification by an Infinity of Conditional Probabilities". In: *Notre Dame Journal of Formal Logic* 50.2, pp. 183–193. DOI: 10.1215/00294527-2009-005.

Audi, Robert (2011). *Epistemology. A Contemporary Introduction to the Theory of Knowledge*. Third Edition. New York: Routledge.

Bacon, Francis (1620/2000). "The New Organon". In: *The New Organon*. Ed. by Jardine, Lisa and Silverthorne, Michael. Cambridge: Cambridge University Press.

Black, Duncan (1986). *The Theory of Committees and Elections*. Dordrecht: Springer.

Black, Max (1954). *Problems of Analysis: Philosophical Essays*. Ithaca: Cornell University Press.

BonJour, Laurence (1988). *The Structure of Empirical Knowledge*. Cambridge: Cambridge University Press.

Bouvère, Karel de (1978). "Synonymous Theories". In: *The theory of models: proceedings of the 1963 International Symposium at Berkeley*. Ed. by Addison, John West, Henkin, León, and Tarski, Alfred. Amsterdam: North-Holland Publishing Co., pp. 402–406.

Bovens, Luc and Hartmann, Stephan (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Brier, Glenn W. (1950). "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78.1, pp. 1–3.

Brössel, Peter and Eder, Anna-Maria A. (2014). "How to Resolve Doxastic Disagreement". In: *Synthese* 191.11, pp. 2359–2381. DOI: 10.1007/s11229-014-0431-4.

Bunn, Derek W. (1981-03). "Two Methodologies for the Linear Combination of Forecasts". In: *Journal of the Operational Research Society* 32.3, pp. 213–222. DOI: 10.1057/jors.1981.44.

Cariani, Fabrizio, Pauly, Marc, and Snyder, Josh (2008). "Decision framing in judgment aggregation". English. In: *Synthese* 163.1, pp. 1–24. DOI: 10.1007/s11229-008-9306-x.

Carnap, Rudolf (1936). "Testability and Meaning". In: *Philosophy of Science* 3.4, pp. 419–471. DOI: 10.1086/286432.

— (1937). "Testability and Meaning - Continued". In: *Philosophy of Science* 4.1, pp. 1–40. DOI: 10.1086/286443.

— (1947). "On the Application of Inductive Logic". In: *Philosophy and Phenomenological Research* 8.1, pp. 133–148. DOI: 10.2307/2102920.

— (1952). *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

— (1959). *Induktive Logik und Wahrscheinlichkeit*. bearbeitet von Wolfgang Stegmüller. Wien: Springer.

— (1966). "The Aim of Inductive Logic". In: *Proceedings of the 1960 International Congress for Logic, Methodology and Philosophy of Science*. Ed. by Nagel, Ernest, Suppes, Patrick, and Tarski, Alfred. Vol. 44. Studies in Logic and the Foundations of Mathematics. Elsevier, pp. 303–318. DOI: 10.1016/S0049-237X(09)70598-1.

— (1950/1962). *Logical Foundations of Probability*. London: Routledge and Kegan Paul.

— (1928/2003). *The Logical Structure of the World and Pseudoproblems in Philosophy*. Open Court Classics. Illinous: Open Court.

Cesa-Bianchi, Nicolo and Lugosi, Gabor (2006). *Prediction, Learning, and Games*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511546921.

Chisholm, Roderick M (1989). *Theory of Knowledge*. Third Edition. Englewood Cliffs: Prentice-Hall International.

Christensen, David (2007). "Epistemology of Disagreement: The Good News". In: *Philosophical Review* 116.2, pp. 187–217. DOI: 10.1215/00318108-2006-035.

Clark, Andy (2002). *Paradoxes from A to Z*. Second Edition. London: Routledge.

Clemen, Robert T. and Winkler, Robert L. (2007). "Aggregating Probability Distributions". In: *Advances in Decision Analysis: From Foundations to Applications*. Ed. by Edwards, Ward, Miles, Ralph F., and Detlof von Winterfeldt. Cambridge: Cambridge University Press, pp. 154–176.

Cohnitz, Daniel, Pagin, Peter, and Rossberg, Marcus (2014-03). “Monism, Pluralism and Relativism: New Essays on the Status of Logic”. In: *Erkenntnis* 79.2, pp. 201–210. DOI: 10.1007/s10670-013-9473-0.

Cohnitz, Daniel and Rossberg, Marcus (2006). *Nelson Goodman*. Chesham: Acumen.

Condorcet, Nicolas, Marquis de (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: L'Imprimerie Royale.

Constantin, Jan and Grundmann, Thomas (2018-09). “Epistemic Authority: Preemption through source sensitive defeat”. In: *Synthese*. DOI: 10.1007/s11229-018-01923-x.

Courgeau, Daniel (2012). *Probability and Social Science: Methodological relationships between the two approaches*. Dordrecht: Springer.

Cover, Thomas (1965). “Behaviour of Sequential Predictors of Binary Sequences”. In: *Transaction of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pp. 263–272.

Dardashti, Radin, Hartmann, Stephan, Thébault, Karim, and Winsberg, Eric (2019). “Hawking Radiation and Analogue Experiments: A Bayesian analysis”. In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*. DOI: https://doi.org/10.1016/j.shpsb.2019.04.004.

Daston, Lorraine (1988). *Classical Probability in the Enlightenment*. Princeton: Princeton University Press.

Descartes, René (1975). *The Philosophical Works of Descartes: Rendered into English, Volume 1*. Ed. by Sanderson Haldane, Elizabeth and Ross, George R. T. Cambridge: Cambridge University Press.

— (1637/1998). *Discourse on Method*. Ed. by Cress, Donald A. Indianapolis: Hackett Publishing Company.

Dietrich, Franz (2008-02). “The Premises of Condorcet's Jury Theorem Are Not Simultaneously Justified”. In: *Episteme* 5.1 (01), pp. 56–73. DOI: 10.3366/E1742360008000233.

Dietrich, Franz and List, Christian (2016). “Probabilistic Opinion Pooling”. In: *The Oxford Handbook of Probability and Philosophy*. Ed. by Hájek, Alan and Hitchcock, Christopher. Oxford: Oxford University Press.

Dietrich, Franz and Spiekermann, Kai (2013). “Epistemic Democracy with Defensible Premises”. In: *Economics and Philosophy* 29.1, pp. 87–120. DOI: 10.1017/S0266267113000096.

Domingos, Pedro (2015). *The Master Algorithm. How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books.

Douven, Igor (2018). “Abduction”. In: *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*. Ed. by Zalta, Edward N.

Douven, Igor and Riegler, Alexander (2010). “Extending the Hegselmann-Krause Model I”. In: *Logic Journal of the IGPL* 18.2, pp. 323–335. DOI: 10.1093/jigpal/jzp059.

Earman, John (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge: The MIT Press.

Easwaran, Kenny, Fenton-Glynn, Luke, Hitchcock, Christopher, and Velasco, Joel D. (2016). "Updating on the Credences of Others: Disagreement, Agreement, and Synergy". In: *Philosophers' Imprint* 16.11, pp. 1–39.

Elga, Adam (2007). "Reflection and Disagreement". English. In: *Noûs* 41.3, pp. 478–502. DOI: 10.1111/j.1468-0068.2007.00656.x.

— (2010). "How to Disagree about how to Disagree". In: *Disagreement*. Ed. by Feldman, Richard and Warfield, Ted A. Oxford: Oxford University Press, pp. 175–186.

Elgin, Catherine Z. (2014). "Non-foundationalist Epistemology: Holism, Coherence, and Tenability". In: *Contemporary Debates in Epistemology*. Ed. by Steup, Matthias, Turri, John, and Sosa, Ernest. 2nd Edition. Chichester: John Wiley & Sons, pp. 244–255.

Estlund, David M. (1994-03). "Opinion Leaders, Independence, and Condorcet's Jury Theorem". In: *Theory and Decision* 36.2, pp. 131–162. DOI: 10.1007/BF01079210.

Fantl, Jeremy (2003). "Modest Infinitism". In: *Canadian Journal of Philosophy* 33.4, pp. 537–562. DOI: 10.1080/00455091.2003.10716554.

Feigl, Herbert (1950). "De Principiis Non Dispudandum ...? On the Meaning and Limits of Justification". In: *Philosophical Analysis*. Ed. by Black, Max. Ithaca: Cornell University Press, pp. 119–156.

— (1981). "De Principiis Non Dispudandum ...? On the Meaning and Limits of Justification [1950]". In: *Inquiries and Provocations. Selected Writings 1929-1974*. Ed. by Cohen, Robert S. Dordrecht: Reidel Publishing Company, pp. 237–268.

Feldbacher-Escamilla, Christian J. (2012). "Meta-Induction and the Wisdom of Crowds. A Comment". In: *Analyse und Kritik* 34.2, pp. 367–382. DOI: 10.1515/auk-2012-0213.

— (2015). "Is the Equal-Weight View Really Supported by Positive Crowd Effects?" In: *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*. Ed. by Mäki, Uskali, Votsis, Ioannis, Ruphy, Stephanie, and Schurz, Gerhard. Heidelberg: Springer, pp. 87–98. DOI: 10.1007/978-3-319-23015-3_7.

— (2017a). "One Dogma of Analyticism". In: *Logique et Analyse* 240, pp. 429–444. DOI: 10.2143/LEA.240.0.3254090.

— (2017b). "Optimisation in a Synchronised Prediction Setting". In: *Journal for General Philosophy of Science* 48.3, pp. 419–437. DOI: 10.1007/s10838-017-9379-7.

— (2019). "Newtons Methodologie: Eine Kritik an Duhem, Feyerabend und Lakatos". In: *Archiv für Geschichte der Philosophie* 101.4, pp. 584–615. DOI: 10.1515/agph-2019-4004.

Feldbacher-Escamilla, Christian J. (2020a). "An Optimality-Argument for Equal Weighting". In: *Synthese* 197.4, pp. 1543–1563. DOI: 10.1007/s11229-018-02028-1.

— (2020b). *Elementare Definitionstheorie*. Stuttgart: Metzler.

— (manuscript). "Abductive Philosophy and Error". In: *manuscript*.

Feldbacher-Escamilla, Christian J. and Gebharter, Alexander (2019-01). "Modeling Creative Abduction Bayesian Style". In: *European Journal for Philosophy of Science* 9.1, pp. 1–15. DOI: 10.1007/s13194-018-0234-4.

Feldbacher-Escamilla, Christian J. and Schurz, Gerhard (2019). "Optimal Probability Aggregation Based on Generalized Brier Scoring". In: *Annals of Mathematics and Artificial Intelligence* forthcoming. DOI: 10.1007/s10472-019-09648-4.

— (minor revisions). "Meta-Inductive Probability Aggregation". In: *manuscript*.

Feldman, Richard (1999). "Methodological Naturalism in Epistemology". In: *The Blackwell Guide to Epistemology*. Ed. by Greco, John and Sosa, Ernest. Malden: Blackwell Publishing, pp. 170–186.

— (2007). "Reasonable Religious Disagreements". In: *Philosophers Without God. Mediation on Atheism and Secular Life*. Ed. by Antony, Louise. Oxford: Oxford University Press, pp. 194–214.

Fieser, James and Dowden, Bradley, eds. (2012). *Internet Encyclopedia of Philosophy*.

Fitelson, Branden (2012). "Evidence of Evidence is Not (Necessarily) Evidence". In: *Analysis* 72.1, pp. 85–88.

Forster, Malcolm R. and Sober, Elliott (1994). "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions". In: *The British Journal for the Philosophy of Science* 45.1, pp. 1–35. DOI: 10.1093/bjps/45.1.1.

Frances, Bryan and Matheson, Jonathan (2018). "Disagreement". In: *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*. Ed. by Zalta, Edward N.

Fumerton, Richard A. (1995). *Metaepistemology and Skepticism*. Lanham: Rowman & Littlefield Publishers, Inc.

— (2002). "Theories of Justification". In: *The Oxford Handbook of Epistemology*. Oxford: Oxford University Press, pp. 204–233.

Gaertner, Wulf C. (2009). *A Primer in Social Choice Theory*. Revised Edition. Oxford: Oxford University Press.

— (2016). "Wickedness in Social Choice". In: *Journal of Economic Surveys*. DOI: 10.1111/joes.12143.

Gallistel, Charles R. (1993). *The Organization of Learning*. Cambridge, Massachusetts: MIT Press.

Galton, Francis (1869). *Hereditary Genius. An inquiry into its laws and consequences*. London: Macmillan and Co.

Galton, Francis (1907a-02). "One Vote, One Value". In: *Nature* 75, p. 414. DOI: 10.1038/075414a0.

— (1907b-03). "The Ballot-Box". In: *Nature* 75, p. 509. DOI: 10.1038/075509f0.

— (1907c-03). "Vox Populi". In: *Nature* 75, pp. 450–451. DOI: 10.1038/075450a0.

— (1908). *Memories of My Life*. London: Methuen & Co.

Gebharter, Alexander (2017). *Causal Nets, Interventionism, and Mechanisms. Philosophical Foundations and Applications*. Cham: Springer.

Genest, Christian (1984-09). "A Characterization Theorem for Externally Bayesian Groups". In: *The Annals of Statistics* 12.3, pp. 1100–1105. DOI: 10.1214/aos/1176346726.

Genest, Christian and McConway, Kevin J. (1990). "Allocating the Weights in the Linear Opinion Pool". In: *Journal of Forecasting* 9.1, pp. 53–73. DOI: 10.1002/for.3980090106.

Genest, Christian, McConway, Kevin J., and Schervish, Mark J. (1986-06). "Characterization of Externally Bayesian Pooling Operators". In: *The Annals of Statistics* 14.2, pp. 487–501. DOI: 10.1214/aos/1176349934.

Genest, Christian and Zidek, James V. (1986-02). "Combining Probability Distributions: A Critique and an Annotated Bibliography". In: *Statistical Sciences* 1.1, pp. 114–135.

Gettier, Edmund L. (1963). "Is Justified True Belief Knowledge?" In: *Analysis* 23.6, pp. 121–123. DOI: 10.1093/analys/23.6.121.

Gillham, Nicholas W. (2001). *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics*. Oxford: Oxford University Press.

Glymour, Clark, Spirtes, Peter, and Scheines, Richard (1991-07). "Causal inference". In: *Erkenntnis* 35.1, pp. 151–189. DOI: 10.1007/BF00388284.

Gold, E. Mark (1967). "Language Identification in the Limit". In: *Information and Control* 10.5, pp. 447–474. DOI: 10.1016/S0019-9958(67)91165-5.

Goldman, Alvin I. (1979). "What Is Justified Belief?" In: *Justification and Knowledge. New Studies in Epistemology*. Ed. by Pappas, George S. Dordrecht: D. Reidel Publishing Company, pp. 1–23.

— (1994). "Naturalistic Epistemology and Reliabilism". In: *Midwest Studies In Philosophy* 19.1, pp. 301–320. DOI: 10.1111/j.1475-4975.1994.tb00291.x.

— (1999). *Knowledge in a Social World*. Oxford: Oxford University Press.

— (2011a). "A Guide to Social Epistemology". In: *Social Epistemology. Essential Readings*. Ed. by Goldman, Alvin I. and Whitcomb, Dennis. Oxford: Oxford University Press, pp. 11–37.

— (2011b). "Experts: Which Ones Should You Trust?" In: *Social Epistemology. Essential Readings*. Ed. by Goldman, Alvin I. and Whitcomb, Dennis. Oxford: Oxford University Press, pp. 109–133.

Goldman, Alvin I. (2014). "Social Process Reliabilism: Solving Justification Problems in Collective Epistemology". In: *Essays in Collective Epistemol-*

ogy. Ed. by Lackey, Jennifer. Oxford: Oxford University Press, pp. 11–41.

Goldman, Alvin I. and Whitcomb, Dennis, eds. (2011). *Social Epistemology. Essential Readings*. Oxford: Oxford University Press.

Goodman, Nelson (1946). "A Query on Confirmation". In: *The Journal of Philosophy* 43.14, pp. 383–385. DOI: 10.2307/2020332.

— (1955/1983). *Fact, Fiction, and Forecast*. Ed. by Putnam, Hilary. Fourth Edition. Harvard: Harvard University Press.

Greco, John and Sosa, Ernest, eds. (1999). *The Blackwell Guide to Epistemology*. Malden: Blackwell Publishing.

Grofman, Bernard, Owen, Guillermo, and Feld, Scott L. (1983-09). "Thirteen theorems in search of the truth". In: *Theory and Decision* 15.3, pp. 261–278. DOI: 10.1007/BF00125672.

Grundmann, Thomas (2009). "Reliabilism and the Problem of Defeaters". In: *Grazer Philosophische Studien* 79.1, pp. 65–76. DOI: 10.1163/18756735-90000857.

— (2017). *Analytische Einführung in die Erkenntnistheorie*. Berlin: de Gruyter.

Haack, Susan (1976). "The Justification of Deduction". In: *Mind* LXXXV.337, pp. 112–119. DOI: 10.1093/mind/LXXXV.337.112.

Hacking, Ian (2006). *The Emergence of Probability. A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Second Edition. Cambridge: Cambridge University Press.

Hájek, Alan (2005). "Scotching Dutch Books?" In: *Philosophical Perspectives* 19.1, pp. 139–151. DOI: 10.1111/j.1520-8583.2005.00057.x.

Hale, R.L.V. (1978). "Logic for Morons". In: *Mind* LXXXVII.1, pp. 111–115. DOI: 10.1093/mind/LXXXVII.1.111.

Hartmann, Stephan (2017-07). *Prospect Theory and the Wisdom of the Inner Crowd*. URL: http://philsci-archive.pitt.edu/13250/.

Hartmann, Stephan, Martini, Carlo, and Sprenger, Jan (2009). "Consensual Decision-Making Among Epistemic Peers". In: *Episteme* 6.2, pp. 110–129. DOI: 10.3366/E1742360009000598.

Hartmann, Stephan and Rafiee Rad, Soroush (2018-03). "Voting, Deliberation and Truth". In: *Synthese* 195.3, pp. 1273–1293. DOI: 10.1007/s11229-016-1268-9.

Hartmann, Stephan and Sprenger, Jan (2010). "The Weight of Competence Under a Realistic Loss Function". In: *Logic Journal of the IGPL* 18.2, pp. 346–352. DOI: 10.1093/jigpal/jzp061.

— (2011). "Bayesian Epistemology". In: *The Routledge Companion to Epistemology*. Ed. by Bernecker, Sven and Pritchard, Duncan. Routledge Philosophy Companions. London: Routledge, pp. 609–620.

— (2012). "Judgment Aggregation and the Problem of Tracking the Truth". In: *Synthese* 187.1, pp. 209–221. DOI: 10.1007/s11229-011-0031-5.

Hempel, Carl G. (1945a). "Studies in the Logic of Confirmation (I.)" In: *Mind* 54.213, pp. 1–26. DOI: 10.1093/mind/LIV.213.1.

— (1945b). “Studies in the Logic of Confirmation (II.)” In: *Mind* 54.214, pp. 97–121. DOI: 10.1093/mind/LIV.214.97.

— (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Henderson, Leah (2018). “The Problem of Induction”. In: *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*. Ed. by Zalta, Edward N.

Herzog, Stefan M. and Hertwig, Ralph (2009). “The Wisdom of Many in One Mind: Improving Individual Judgments With Dialectical Bootstrapping”. In: *Psychological Science* 20.2. PMID: 19170937, pp. 231–237. DOI: 10.1111/j.1467-9280.2009.02271.x.

Hitchcock, Christopher and Sober, Elliott (2004). “Prediction Versus Accommodation and the Risk of Overfitting”. In: *The British Journal for the Philosophy of Science* 55.1, pp. 1–34. DOI: 10.1093/bjps/55.1.1.

Hodge, Victoria and Austin, Jim (2004-10). “A Survey of Outlier Detection Methodologies”. In: *Artificial Intelligence Review* 22.2, pp. 85–126. DOI: 10.1023/B:AIRE.0000045502.10941.a9.

Howson, Colin (2000). *Hume’s Problem*. Oxford: Clarendon Press.

Howson, Colin and Urbach, Peter (2006). *Scientific Reasoning: The Bayesian Approach*. Third Edition. La Salle, Illinois: Open Court.

Huber, Franz (2017-10). “On the Justification of Deduction and Induction”. In: *European Journal for Philosophy of Science* 7.3, pp. 507–534. DOI: 10.1007/s13194-017-0177-1.

Hume, David (1772). *Essays and Treatises on Several Subjects. Vol. II: An Enquiry Concerning Human Understanding, A Dissertation on the Passions, An Enquiry Concerning the Principles of Morals, and The Natural History of Religion*. London: T. Cadell.

— (1738/1960). *A Treatise of Human Nature*. Ed. by Selby-Bigge, L.A. Oxford: Clarendon Press.

— ed. (1748/2007). *An Enquiry Concerning Human Understanding*. Oxford World’s Classics. (ed. by Millican, Peter). Oxford University Press.

Jacquette, Dale (2011). “How (Not) to Justify Induction”. In: *Kriterion – Journal of Philosophy* 24.1, pp. 01–18.

Kahneman, Daniel and Tversky, Amos (1979). “Prospect Theory: Analysis of Decision under Risk”. In: *Econometrica* 47.2, pp. 263–291.

Kanger, Stig (1968). “Equivalent Theories”. In: *Theoria* 34, pp. 1–6. DOI: 10.1111/j.1755;%20--2567.1968.

Kant, Immanuel (1787/1998). *Critique of Pure Reason*. Ed. by Guyer, Paul and Wood, Allen W. Translated by the editors. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

— (1762/1997). *Lectures on Ethics*. Ed. by Heath, Peter and Schneewind, J.B. Translated by Peter Heath. Cambridge: Cambridge University Press.

Kelly, Kevin T. (1996). *The Logic of Reliable Inquiry*. Oxford: Oxford University Press.

Kelly, Kevin T., Schulte, Oliver, and Juhl, Cory (1997). “Learning Theory and the Philosophy of Science”. In: *Philosophy of Science* 64.2, pp. 245–267. DOI: 10.1086/392550.

Kelly, Thomas (2011). “Peer Disagreement and Higher Order Evidence”. In: *Social Epistemology. Essential Readings*. Ed. by Goldman, Alvin I. and Whitcomb, Dennis. Oxford: Oxford University Press, pp. 183–217.

Keren, Arnon (2014). “Zagzebski on Authority and Preemption in the Domain of Belief”. In: *European Journal for Philosophy of Religion* 6.4, pp. 61–76.

Keynes, John Maynard (1921). *A Treatise on Probability*. London: Macmillan and Co.

Klein, Peter D. (1998). “Foundationalism and the Infinite Regress of Reasons”. In: *Philosophy and Phenomenological Research* 58.4, pp. 919–925. DOI: 10.2307/2653735.

— (2014). “Infinitism Is the Solution to the Regress Problem”. In: *Contemporary Debates in Epistemology*. Ed. by Steup, Matthias, Turri, John, and Sosa, Ernest. 2nd Edition. Chichester: John Wiley & Sons, pp. 274–283.

Klein, Peter D. and Turri, John (2012). “Infinitism in Epistemology”. In: *Internet Encyclopedia of Philosophy*. Ed. by Fieser, James and Dowden, Bradley. URL: http://www.iep.utm.edu/inf-epis/.

Kolmogorov, Andrey N. (1956). *Foundations of the Theory of Probability*. New York: Chelsea Publishing Company.

Kornblith, Hilary (1999). “In Defense of a Naturalized Epistemology”. In: *The Blackwell Guide to Epistemology*. Ed. by Greco, John and Sosa, Ernest. Malden: Blackwell Publishing, pp. 158–169.

— (2002). *Knowledge and its Place in Nature*. Oxford: Clarendon Press.

Kornhauser, Lewis A. (1992). “Modeling Collegial Courts”. In: *Journal of Law, Economics, & Organization* 8.3, pp. 441–470.

Krogh, Anders and Vedelsby, Jespers (1995). “Neural Network Ensembles, Cross Validation, and Active Learning”. In: *Advances in Neural Information Processing Systems 7*. Ed. by Tesauro, Gerald, Touretzky, David, and Leen, Todd. Cambridge: The MIT Press, pp. 231–238.

Kyburg (Jr.), Henry and Teng, Choh Man (2001). *The Theory of Probability*. Cambridge: Cambridge University Press.

Lackey, Jennifer (2008). *Learning From Words. Testimony as a Source of Knowledge*. Oxford: Oxford University Press.

— (2011). “Testimony: Acquiring Knowledge from Others”. In: *Social Epistemology. Essential Readings*. Ed. by Goldman, Alvin I. and Whitcomb, Dennis. Oxford: Oxford University Press, pp. 71–91.

Lakatos, Imre (1970). “Falsification and the Methodology of Scientific Research Programmes”. In: *Criticism and the Growth of Knowledge*. Ed. by Lakatos, Imre and Musgrave, Alan. Cambridge: Cambridge University Press, pp. 91–196.

Larrick, Richard P. and Soll, Jack B. (2006). "Intuitions About Combining Opinions: Misappreciation of the Averaging Principle". In: *Management Science* 52.1, pp. 111–127. DOI: 10.1287/mnsc.1050.0459.

Laudan, Larry (1981). *Science and Hypothesis. Historical Essays on Scientific Methodology*. Dordrecht: Springer Science + Business Media.

Lehrer, Keith and Wagner, Carl (1981). *Rational Consesus in Science and Society. A Philosophical and Mathematical Study*. Dordrecht: Reidel Publishing Company.

— (1983). "Probability Amalgamation and the Independence Issue: A reply to Laddaga". In: *Synthese* 55.3, pp. 339–346. DOI: 10.1007/BF00485827.

Leibniz, Gottfried Wilhelm, ed. (1989a). *Philosophical Essays*. Translated by the editors. (ed. Ariew, Roger and Garber, Daniel). Indianapolis: Hackett Publishing Company.

— (1989b). "Preface to the New Essays (1703-5)". In: *Philosophical Essays*. Ed. by Leibniz, Gottfried Wilhelm. Translated by the editors. (ed. Ariew, Roger and Garber, Daniel). Indianapolis: Hackett Publishing Company, pp. 291–305.

Leitgeb, Hannes (2016). "Imaging All the People". In: *Episteme*, pp. 1–17. DOI: 10.1017/epi.2016.14.

— (2017). *The Stability of Belief. How Rational Belief Coheres with Probability*. Oxford: Oxford University Press.

Lewis, CI (1946). *An Analysis of Knowledge and Valuation*. La Salle: Open Court.

Lipton, Peter (2004). *Inference to the Best Explanation*. 2nd Edition. London: Routledge.

List, Christian (2011). "Group Knowledge and Group Rationality: A Judgment Aggregation Perspective". In: *Social Epistemology. Essential Readings*. Ed. by Goldman, Alvin I. and Whitcomb, Dennis. Oxford: Oxford University Press, pp. 221–241.

List, Christian and Pettit, Philip (2002). "Aggregating Sets of Judgments: An Impossibility Result". In: *Economics and Philosophy* 18.01, pp. 89–110.

— (2011). *Group Agency. The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Littlestone, Nick (1988-04). "Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm". In: *Machine Learning* 2.4, pp. 285–318. DOI: 10.1023/A:1022869011914.

Locke, John (1690/1999). *An Essay Concerning Human Understanding*. Ed. by Nidditch, Peter H. Oxford: Clarendon Press.

Longino, Helen E. (2008). "Values, Heuristics, and the Politics of Knowledge". In: *The Challenge of the Social and the Pressure of Practice: Science and Values Revisited*. Ed. by Carrier, Martin, Howard, Don, and Kourany, Janet A. Pittsburgh: University of Pittsburgh Press, pp. 68–87.

Lyon, Aidan (forthcoming). "Collective Wisdom". In: *The Journal of Philosophy*.

Mackay, Charles (1852). *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds*. London: Office of the National Illustrated Library.

Magnani, Lorenzo (2000). *Abduction, Reason and Science. Processes of Discovery and Explanation*. New York: Kluwer Academic Publishers.

Mill, John St. (1843/1974). *A System of Logic. Ratiocinative and Inductive. Part I (The Collected Works of John Stuart Mill - Volume 07)*. Toronto: University of Toronto Press.

Miller, David (1974). "Popper's Qualitative Theory of Verisimilitude". In: *The British Journal for the Philosophy of Science* 25.2, pp. 166–177.

— (2006). *Out of error: Further essays on critical rationalism*. Hampshire: Ashgate Publishing.

Mitchell, Tom (1997). *Machine Learning*. New York: McGraw-Hill.

Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet (2012). *Foundations of Machine Learning*. Cambridge, Massachusetts: The MIT Press.

Mongin, Philippe (2001). "The Paradox of the Bayesian Experts". In: *Foundations of Bayesianism*. Ed. by Corfield, David and Williamson, Jon. Dordrecht: Springer Science+Business Media, pp. 309–338.

Myrvold, Wayne C. (2017). "On the Evidential Import of Unification". In: *Philosophy of Science* 84.1, pp. 92–114.

Neurath, Otto (1932). "Protokollsätze". In: *Erkenntnis* 3.1, pp. 204–214. DOI: 10.1007/BF01886420.

— (2006). "Protokollsätze (1932)". In: *Wiener Kreis. Texte zur wissenschaftlichen Weltauffassung von Rudolf Carnap, Otto Neurath, Moritz Schlick, Philipp Frank, Hans Hahn, Karl Menger, Edgar Zilsel und Gustav Bermann*. Ed. by Stöltzner, Michael and Uebel, Thomas. Hamburg: Felix Meiner, pp. 399–411.

Newton, Isaac (1934). *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and his System of World*. Translated into English by Andrew Motte in 1729. The translations revised, and supplied with an historical and explanatory appendix, by Florian Cajori. Berkely: University of California Press.

Niiniluoto, Ilkka (1987). *Truthlikeness*. illustrated. Synthese library (Vol 185). Berlin: Springer.

— (1998). "Verisimilitude: The Third Period". In: *The British Journal for the Philosophy of Science* 49.1, pp. 1–29. DOI: 10.1093/bjps/49.1.1.

— (1999). "Defending Abduction". In: *Philosophy of Science* 66, S436–S451. DOI: 10.1086/392744.

Norton, John D. (2014-03). "A material dissolution of the problem of induction". In: *Synthese* 191.4, pp. 671–690. DOI: 10.1007/s11229-013-0356-3.

Oddie, Graham (2013-06). “The Content, Consequence and Likeness Approaches to Verisimilitude: compatibility, trivialization, and underdetermination”. In: *Synthese* 190.9, pp. 1647–1687. DOI: 10.1007/s11229-011-9930-8.

Olsson, Erik J. (2018). “Coherentist Theories of Epistemic Justification”. In: *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*. Ed. by Zalta, Edward N. URL: https://plato.stanford.edu/archives/spr2017/entries/justep-coherence/.

Osherson, Daniel N., Stob, Michael, and Weinstein, Scott (1986). *Systems That Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists*. Cambridge, Massachusetts: MIT Press.

Page, Scott E. (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton: Princeton University Press.

Pearl, Judea (2000). *Causality. Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearson, Karl (1924). *The Life, Letters and Labours of Francis Galton. Volume Two: Researches of Middle Life*. Cambridge: Cambridge University Press.

Peirce, Charles S. (1994). “Pragmatism and Abduction”. In: *Collected Papers of Charles Sanders Peirce*. Ed. by Hartshorne, Charles, Weiss, Paul, and Burks, Arthur W. Harvard: Harvard University Press.

Pelletier, Francis Jeffry and Urquhart, Alasdair (2003). “Synonymous Logics”. In: *Journal of Philosophical Logic* 32, pp. 259–285.

Pivato, Marcus (2008). “The Discursive Dilemma and Probabilistic Judgement Aggregation”. In: *MPRA. Munich Personal RePEc Archive* 8412. URL: http://mpra.ub.uni-muenchen.de/8412.

Plato, ed. (1997). *Complete Works*. (ed. by Cooper, John M.) Indianapolis: Hackett Publishing Company.

Pojman, Louis P. (2000). *What Can We Know? An Introduction to the Theory of Knowledge*. Belmont: Wadsworth.

Popper, Karl R. (1972). *Objective Knowledge*. Oxford: Oxford University Press.

— (2002a). *Conjectures and Refutation*. New York: Basic Books.

— (2002b). *The Logic of Scientific Discovery*. London: Routledge.

Pritchard, Duncan (2007). “The Value of Knowledge”. In: *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*. Ed. by Zalta, Edward N.

Putnam, Hilary (1963). “Degree of Confirmation and Inductive Logic”. In: *The Philosophy of Rudolf Carnap*. Ed. by Schilpp, Paul Arthur. La Salle: Open Court, pp. 761–784.

— ed. (1981). *Reason, Truth And History*. Cambridge: Cambridge University Press.

Quine, Willard van Orman, ed. (1963a). *From a Logical Point of View. 9 Logico-Philosophical Essays*. New York: Harper Torchbooks.

— (1963b). “Two Dogmas of Empiricism”. In: *From a Logical Point of View. 9 Logico-Philosophical Essays*. Ed. by Quine, Willard van Orman. New York: Harper Torchbooks, pp. 20–46.

The Ways of Paradox and Other Essays (1966). New York: Random House.

Quine, Willard van Orman, ed. (1969). *Ontological Relativity and Other Essays*. Cambridge, Massachusetts: Columbia Universiy Press.

— (1998). “Reply to Morton White”. In: *The Philosophy of W.V. Quine*. Ed. by Hahn, Lewis and Schilpp, Paul Arthur. La Salle: Open Court, pp. 663–665.

Quine, Willard van Orman and Ullian, Joe S. (1978). *The Web of Belief*. Second Edition. New York: McGraw-Hill.

Rakhlin, Alexander, Sridharan, Karthik, and Tewari, Ambuj (2010). “Online Learning: Random Averages, Combinatorial Parameters, and Learnability”. In: *Advances in Neural Information Processing Systems 23*. Ed. by Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. New York: Curran Associates, Inc., pp. 1984–1992. URL: http://papers.nips.cc/paper/4116-online-learning-random-averages-combinatorial-parameters-and-learnability.pdf.

Raz, Joseph (1988). *The Morality of Freedom*. Oxford: Oxford University Press.

Reichenbach, Hans (1938). *Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge*. Chicago: University of Chicago Press.

— (1940). “On the Justification of Induction”. In: *The Journal of Philosophy* 37.4, pp. 97–103. DOI: 10.2307/2017409.

— (1971). *The Direction of Time*. Ed. by Reichenbach, Maria. Berkeley: University of California Press.

Reid, Thomas (1764/1785/1788/1983). *Inquiry and Essays*. Ed. by Beanblossom, Ronald E. and Lehrer, Keith. Indianapolis: Hackett Publishing Company.

Rosen, Gideon (2001). “Nominalism, Naturalism, Epistemic Relativism”. In: *Noûs* 35, pp. 69–91. DOI: 10.1111/0029-4624.35.s15.4.

Rumelhart, Donald L. and Greeno, James G. (1971). “Similarity Between Stimuli: An Experimental test of the Luce and Restle Choice Models”. In: *Journal of Mathematical Psychology* 8, pp. 370–381.

Russell, Jeffrey Sanford, Hawthorne, John, and Buchak, Lara (2015-05). “Groupthink”. In: *Philosophical Studies* 172.5, pp. 1287–1309. DOI: 10.1007/s11098-014-0350-8.

Rysiew, Patrick (2018). “Naturalism in Epistemology”. In: *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*. Ed. by Zalta, Edward N. URL: https://plato.stanford.edu/archives/spr2017/entries/epistemology-naturalized/.

Salmon, Wesley (1963). “On Vindicating Induction”. In: *Philosophy of Science* 30.3, pp. 252–261. DOI: 10.1086/287939.

Salmon, Wesley C. (1957). "Should We Attempt to Justify Induction?" In: *Philosophical Studies* 8.3, pp. 33–48. DOI: 10.1007/BF02308902.

Savage, Leonard J. (1972). *The Foundations of Statistics*. Second Revised Edition. New York: Dover Publications.

Schapire, Robert E. (2012). *Boosting. Foundations and Algorithms*. Ed. by Freund, Yoav. Cambridge, Massachusetts: The MIT Press.

Schupbach, Jonah N. (2005). "On a Bayesian Analysis of the Virtue of Unification". In: *Philosophy of Science* 72.4, pp. 594–607. DOI: 10.1086/505186.

Schurz, Gerhard (1997). *The Is-Ought Problem. An Investigation in Philosophical Logic*. Dordrecht: Kluwer Academic Publishers.

— (2004). "Meta-Induction and the Prediction Game: A New View on Hume's Problem". In: *Knowledge and Belief. Wissen und Glauben*. Ed. by Löffler, Winfried and Weingartner, Paul. Vienna: öbv & hpt, pp. 244–255.

— (2008a). "Patterns of Abduction". English. In: *Synthese* 164.2, pp. 201–234. DOI: 10.1007/s11229-007-9223-4.

— (2008b). "The Meta-Inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem". In: *Philosophy of Science* 75.3, pp. 278–305. DOI: 10.1086/592550.

— (2009). "Meta-Induction and Social Epistemology: Computer Simulations of Prediction Games". In: *Episteme* 6.02, pp. 200–220. DOI: 10.3366/E1742360009000641.

— (2012a). "Meta-Induction and the Problem of Fundamental Disagreement". In: *Epistemology: Contexts, Values, Disagreement. Proceedings of the 34th International Ludwig Wittgenstein-Symposium in Kirchberg am Wechsel, Austria 2011*. Ed. by Jäger, Christoph and Löffler, Winfried. Frankfurt: Ontos, pp. 343–354.

— (2012b). "Meta-Induction in Epistemic Networks and the Social Spread of Knowledge". In: *Episteme* 9 (Special Issue 02), pp. 151–170. DOI: 10.1017/epi.2012.6.

— (2013). *Philosophy of Science. A Unified Approach*. New York: Routledge.

— (2014). "Bayesian Pseudo-Confirmation, Use-Novelty, and Genuine Confirmation". In: *Studies in History and Philosophy of Science Part A* 45, pp. 87–96. DOI: 10.1016/j.shpsa.2013.10.008.

— (2016). "Common Cause Abduction: The formation of theoretical concepts and models in science". In: *Logic Journal of the IGPL* 24.4, pp. 494–509. DOI: 10.1093/jigpal/jzw029.

— (2017-12). "No Free Lunch Theorem, Inductive Skepticism, and the Optimality of Meta-Induction". In: *Philosophy of Science* 84.5, pp. 825–839. DOI: 10.1086/693929.

Schurz, Gerhard (2018-03). "Optimality justifications: new foundations for foundation-oriented epistemology". In: *Synthese*. DOI: 10.1007/s11229-017-1363-6.

— (2019). *Hume's Problem Solved. The Optimality of Meta-Induction*. Cambridge, Massachusetts: The MIT Press.

Schurz, Gerhard and Gebharter, Alexander (2016-04). "Causality as a Theoretical Concept: explanatory warrant and empirical content of the theory of causal nets". In: *Synthese* 193.4, pp. 1073–1103. DOI: 10.1007/s11229-014-0630-z.

Schurz, Gerhard and Thorn, Paul D. (2016). "The Revenge of Ecological Rationality: Strategy-Selection by Meta-Induction Within Changing Environments". In: *Minds and Machines* 26.1, pp. 31–59. DOI: 10.1007/s11023-015-9369-7.

Schurz, Gerhard and Weingartner, Paul (1987). "Verisimilitude Defined by Relevant Consequence-Elements. A New Reconstruction of Popper's Original Idea". In: *What is Closer-to-the-Truth? A parade of approaches to truthlikeness*. Ed. by Kuipers, Theo. Amsterdam: Rodopi, pp. 47–77.

— (2010). "Zwart and Franssen's impossibility theorem holds for possible-world-accounts but not for consequence-accounts to verisimilitude". In: *Synthese* 172 (3). 10.1007/s11229-008-9399-2, pp. 415–436. DOI: 10.1007/s11229-008-9399-2.

Sellars, Wilfrid (1991). *Science, Perception and Reality*. Ridgeview Pub Co.

Sextus Empiricus (1999). "Outlines of Pyrrhonism, Book I, Sections 1-16, 18-27". In: *Epistemology: The Classic Readings*. Ed. by Cooper, David E. Oxford: Blackwell Publishers Ltd, pp. 43–59.

Shalev-Shwartz, Shai and Ben-David, Shai (2014). *Understanding Machine Learning. From Theory to Algorithms*. Cambridge: Cambridge University Press.

Skyrms, Brian (2000). *Choice and Chance. An Introduction to Inductive Logic*. Fourth Edition. Stamford: Wadsworth.

Solomonoff, Ray J. (1964). "A Formal Theory of Inductive Inference. Part I". In: *Information and Control* 7.1, pp. 1–22. DOI: 10.1016/S0019-9958(64)90223-2.

Sorensen, Roy (2018). "Epistemic Paradoxes". In: *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*. Ed. by Zalta, Edward N. URL: https://plato.stanford.edu/archives/sum2018/entries/epistemic-paradoxes/.

Spirtes, Peter (2009). "Variable Definition and Causal Inference". In: *Logic, Methodology and Philosophy of Science. Proceedings of the Thirteenth International Congress*. Ed. by Glymour, Clark, Wei, Wang, and Westerstahl, Dag. London: College Publications: King's College London, pp. 514–537.

Spirtes, Peter, Glymour, Clark, and Scheines, Richard (2000). *Causation, Prediction, and Search*. Cambridge, Massachusetts: The MIT Press.

Spohn, Wolfgang (2012). *The Laws of Belief: Ranking Theory and its Philosophical Applications*. Oxford: Oxford University Press.

Sprenger, Jan (2016). "Confirmation and Induction". In: *The Oxford Handbook of Philosophy of Science*. Ed. by Humphreys, Paul. Oxford: Oxford University Press, pp. 185–209. URL: http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199368815.001.0001/oxfordhb-9780199368815-e-10.

Sprenger, Jan and Hartmann, Stephan (2019). *Bayesian Philosophy of Science. Variations on a Theme by the Reverend Thomas Bayes*. Oxford: Oxford University Press.

Sterkenburg, Tom F. (2018-01). "Universal Prediction". PhD dissertation. PhD thesis. Groningen: University of Groningen. URL: http://philsci-archive.pitt.edu/14186/.

— (2019). "The Meta-Inductive Justification of Induction". In: *Episteme*, pp. 1–23. DOI: 10.1017/epi.2018.52.

Steup, Matthias, Turri, John, and Sosa, Ernest, eds. (2014). *Contemporary Debates in Epistemology*. 2nd Edition. Chichester: John Wiley & Sons.

Surowiecki, James (2005). *The Wisdom of Crowds*. New York: Anchor Books.

Talbott, William J. (2008). "Bayesian Epistemology [online]". In: *The Stanford Encyclopedia of Philosophy (Winter 2011 Edition)*. Ed. by Zalta, Edward N. URL: http://plato.stanford.edu/archives/win2011/entires/epistemology-bayesian.

Tarski, Alfred (1936). "Der Wahrheitsbegriff in den formalisierten Sprachen". In: *Studia Philosophica* 1, pp. 261–405.

Thomas Aquinas (1981). *Summa Theologica*. Christian Classics.

Thorn, Paul D. (2018). "Induction by Direct Inference Meets the Goodman Problem". In: *Kriterion – Journal of Philosophy* 32.2, pp. 1–24.

Thorn, Paul D. and Schurz, Gerhard (2012). "Meta-Induction and the Wisdom of Crowds". In: *Analyse und Kritik* 34.2, pp. 339–366.

— (2016). "Attractivity Weighting: Take-the-Best's Foolproof Sibling". In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society, Cognitive Science Society 2016*. Ed. by Papafragou, A., Grodner, D., Mirman, D., and Trueswell, J. C., pp. 456–461.

— (forthcoming). "Meta-Inductive Prediction based on Attractivity Weighting: An Empirical Performance Evaluation". In: *manuscript*.

Thorsrud, Harald (2009). *Ancient Scepticism*. Stocksfield: Acumen.

— (2017). "Ancient Greek Skepticism". In: *Internet Encyclopedia of Philosophy*. Ed. by Fieser, James and Dowden, Bradley.

Tichý, Pavel (1974). "On Popper's Definitions of Verisimilitude". In: *The British Journal for the Philosophy of Science* 25.2, pp. 155–160. DOI: 10.1093/bjps/25.2.155.

— (1976). "Verisimilitude Redefined". In: *The British Journal for the Philosophy of Science* 27.1, pp. 25–42. DOI: 10.1093/bjps/27.1.25.

Turri, John (2009). "On the Regress Argument for Infinitism". In: *Synthese* 166.1, pp. 157–163. DOI: 10.1007/s11229-007-9270-x.

Van Cleve, James (2014). “Why Coherence Is Not Enough: A Defense of Moderate Foundationalism”. In: *Contemporary Debates in Epistemology*. Ed. by Steup, Matthias, Turri, John, and Sosa, Ernest. 2nd Edition. Chichester: John Wiley & Sons, pp. 255–267.

Verein Ernst Mach (1996). “The Scientific Conception of the World. The Vienna Circle”. In: *The Emergence of Logical Empiricism: from 1900 to the Vienna Circle*. Ed. by Sarkar, Sahorta. New York: Garland Publishing, pp. 321–340.

Verschuur, Gerrit L. (1993). *Hidden Attraction: The Mystery and History of Magnetism*. Oxford: Oxford University Press.

Vickers, John (2010). “The Problem of Induction”. In: *The Stanford Encyclopedia of Philosophy (Winter 2010 Edition)*. Ed. by Zalta, Edward N. URL: http://plato.stanford.edu/archives/win2010/entries/induction-problem.

Vineberg, Susan (2016). “Dutch Book Arguments”. In: *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*. Ed. by Zalta, Edward N.

Vul, Edward and Pashler, Harold (2008). “Measuring the Crowd Within: Probabilistic Representations Within Individuals”. In: *Psychological Science* 19.7. PMID: 18727777, pp. 645–647. DOI: 10.1111/j.1467-9280.2008.02136.x.

Vulkan, Nir (2000). “An Economist’s Perspective on Probability Matching”. In: *Journal of Economic Surveys* 14.1, pp. 101–118. DOI: 10.1111/1467-6419.00106.

Williamson, Timothy (2016). “Abductive Philosophy”. In: *The Philosophical Forum* 47.3-4, pp. 263–280. DOI: 10.1111/phil.12122.

Wittgenstein, Ludwig (1961). *Tractatus Logico-Philosophicus*. Translation by D.F. Pears & B.F. McGuinnes. London: Routledge and Kegan Paul.

Wolpert, David H. (1996). “The Lack of A Priori Distinctions Between Learning Algorithms”. In: *Neural Computation* 8.7, pp. 1341–1390. DOI: 10.1162/neco.1996.8.7.1341.

Wolpert, David H. and Macready, William G. (1997). “No Free Lunch Theorems for Optimization”. In: *IEEE Transactions on Evolutionary Computation* 1.1, pp. 67–82.

Woodward, James (2018). “Scientific Explanation”. In: *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*. Ed. by Zalta, Edward N.

Zagzebski, Linda (2012). *Epistemic Authority. A Theory of Trust, Authority, and Autonomy in Belief*. Oxford: Oxford University Press.

Zalta, Edward N., ed. (2010). *The Stanford Encyclopedia of Philosophy (Winter 2010 Edition)*.

— ed. (2018). *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*.

Zollman, Kevin J.S. (2007). “The Communication Structure of Epistemic Communities”. In: *Philosophy of Science* 74.5, pp. 574–587. DOI: 10.1086/525605.

— (2014). “A Systems-Oriented Approach to the Problem of Testimony”. In: manuscript.