THE META-INDUCTIVE JUSTIFICATION OF INDUCTION

TOM F. STERKENBURG

Abstract. I evaluate Schurz's proposed meta-inductive justification of induction, a refinement of Reichenbach's pragmatic justification that rests on results from the machine learning branch of prediction with expert advice.

My conclusion is that the argument, suitably explicated, comes remarkably close to its grand aim: an actual justification of induction. This finding, however, is subject to two main qualifications, and still disregards one important challenge.

The first qualification concerns the empirical success of induction. Even though, I argue, Schurz's argument does not need to spell out what inductive method actually consists in, it does need to postulate that there is something like the inductive or scientific prediction strategy that has so far been *significantly* more successful than alternative approaches. The second qualification concerns the difference between having a justification for inductive method and for sticking with induction *for now*. Schurz's argument can only provide the latter. Finally, the remaining challenge concerns the pool of alternative strategies, and the relevant notion of a meta-inductivist's optimality that features in the analytical step of Schurz's argument. Building on the work done here, I will argue in a follow-up paper that the argument needs a stronger *dynamic* notion of a meta-inductivist's optimality.

1. Introduction

Schurz (2008; 2009; most recently, 2018; 20xx) proposes a justification of induction based on the provable optimality of *meta-induction*, induction at the level of competing methods of inference.

1.1. Schurz's proposal.

The naive reply to the problem of induction is that induction is justified because it has been successful in the past. Naive, Hume showed, because this suggested justification must be circular: it requires the principle of induction itself. Either that, one can qualify, or it must rest on a higher principle of induction: the justification of *object*-induction at the level of observed events requires *meta*-induction at the level of inference methods (see Skyrms, 2000, 35ff).

Date: December 19, 2018.

This is the final version, as accepted for publication in *Episteme* (doi: 10.1017/epi.2018.52). Part of the research for this paper was done while I was with the Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam (CWI) and the Faculty of Philosophy, University of Groningen; and while I was a visiting fellow at the Tilburg Center for Logic, Ethics, and Philosophy of Science (TiLPS). For valuable discussion and feedback I would like to thank, especially, Gerhard Schurz. Thanks also to the participants of a seminar meeting at the TiLPS on this work, including Seamus Bradley, Colin Elliot, Felipe Romero, Sander Verhaegh, and Naftali Weinberger; and the participants of a meeting on an early version of this paper at UC Irvine, Daniel Herrmann, Simon Huttegger, and Brian Skyrms. Further thanks go to Tim van Erven, Christian Feldbacher-Escamilla, Peter Grünwald, Stephan Hartmann, Leah Henderson, Wouter Koolen, Jan-Willem Romeijn, Jan Sprenger, Marta Sznajder, and Paul Thorn.

But then induction at *this* level would require justification, setting in motion an infinite regress.

Now Schurz's idea is essentially that this regress may be halted directly at the second level: his idea is that it is possible to give a purely *deductive* or *analytical* justification of meta-induction, induction at the level of methods. This deductive justification comes in the form of mathematical results on the optimality of meta-inductive strategies in the setting of sequential prediction, results from the field of *prediction with expert advice* (Littlestone and Warmuth, 1994; Vovk, 1990, 1998; Cesa-Bianchi et al., 1997; for overviews see Cesa-Bianchi and Lugosi, 2006; Vovk, 2001; Grünwald, 2007, 573ff).

Furthermore, "this analytic justification of *meta-induction* would at the same time yield an *a posteriori* justification of *object-induction* in the real world: for we know by experience that in the real world, noninductive prediction strategies have not been successful so far, hence it would be meta-inductively justified to favor object-inductivistic strategies" (Schurz, 2008, 282).

Schurz's proposal is a refinement of the basic idea behind Reichenbach's attempted *vindication* of induction (Reichenbach, 1933, 421f; 1935, 479ff; 1938, 348ff; Feigl, 1950; see Salmon, 1967, 52ff, 85ff; Skyrms, 2000, 44ff). This idea is that, while we cannot deductively show induction to be *reliable* or guaranteed to be successful, without the help of question-begging assumptions on the 'uniformity of nature,' we might be able to show deductively that induction is *optimal*: it is guaranteed to be successful, *no matter* the course of nature, *if any method is*.

1.2. **Reichenbach's pragmatic justification.** Reichenbach's proposal was based on a particular rule of induction, that was motivated by his thoroughly probabilistic epistemology and his frequentist interpretation of probability. This induction rule, best known as the *straight rule*, infers probabilities of future events through induction by enumeration: it estimates the limiting relative frequencies by the *current* relative frequencies.

Now Reichenbach's argument is as follows. Either nature is in fact *uniform*—a limiting relative frequency exists—or it is not. If it is, then the inductive method (the straight rule) will be successful: it will converge on the correct limiting frequency. If it is not, if no limiting relative frequency exists, then obviously *no* method will be successful in this sense. Hence the inductive method is successful if *any* strategy is.

Clearly, there are many weak spots in this argument. There is the sweeping reduction of scientific inference to estimating limiting relative frequencies (Sellars, 1964, 212ff; Skyrms, 1965, 254ff); there is the fact that there are infinitely many other methods that are likewise guaranteed to be successful in converging to an existing limiting relative frequency (Reichenbach, 1938; Salmon, 1957). But most destructive to the argument is the fact that *it is simply not true* that no method can be successful if there is no limiting relative frequency (Herz, 1936). After all, event sequences that never converge on a particular relative frequency of events might still be successfully—even perfectly—predicted by one or another method: because such sequences might still have much structure, or indeed because we cannot exclude the 'method' of a clairvoyant with perfect foresight.

To this last objection Reichenbach (1938, 357ff) replied that a successful such alternative method *has a high relative frequency of successful predictions*, which implies that his inductive method posits that this method's predictions will continue

to be accurate (by enumerative induction the *limiting* relative frequency of this method's predictions being successful is inferred to be high). Reichenbach did not proceed to make this idea more precise, though: and indeed it does not seem feasible to reconcile the different levels at which his straight rule is now working—both the *object*-level of the events and the *meta*-level of other methods—in such a way that his strategy can be vindicated as desired (Skyrms, 1965, 260f; 2000, 44ff; Schurz, 2008, 281).

1.3. **Schurz's proposal; this paper.** By clearly separating object-induction on the level of events and meta-induction on the level of prediction strategies, Schurz manages to give a precise form to Reichenbach's fundamental observation that induction picks up the past success of alternative methods. By combining this with the *empirical* fact of the object-inductive strategy's success, Schurz claims to have derived a proper, noncircular justification of object-induction.

Can Schurz's proposal be considered successful? In this paper I work towards an answer by reconstructing the argument, and clarifying and filling in aspects that, I think, are somewhat opaque or missing in Schurz's own presentation. My conclusion will be that the argument, suitably explicated, goes a long way; but there are some qualifications. One qualification is the need for a very strong empirical postulate on the success of the *scientific method*, that, I will claim, we must identify the object-inductive strategy with. Another is that the argument cannot provide justification for the object-inductive *strategy*; at best it can give justification for following the object-inductivist, or the scientific method, *for now*. I also identify a subtle yet important issue having to do with the relevant pool of alternative strategies, that I leave for further investigation in a follow-up paper.

Another aim of this paper is to provide a brief and accessible overview of some main results and insights from prediction with expert advice. While this approach is as of yet little known outside of the theory of machine learning, its core ideas, particularly the optimality of aggregating experts' predictions based on their past predictive success, promise to offer an interesting new perspective on *opinion pooling* in social epistemology (see Dietrich and List, 2016). In particular, I will embed these results in a more familiar *Bayesian* perspective.

1.4. **Overview of this paper.** First, in section 2, I will discuss the framework of sequential prediction, including the notion of prediction game (the specification of possible events and predictions, sect. 2.1, and a loss function, sect. 2.2), the goal of designing an optimal strategy (sect. 2.3), and its adequacy for studying the problem of induction (sect. 2.4). I will give special attention to the *probabilistic* prediction game.

Next, in section 3, I will give a detailed reconstruction of Schurz's argument. The first step (sect. 3.1) is the analytical optimality of meta-induction: here I will explain the relevant definitions of meta-inductive strategies and the relevant results regarding their optimality, and I will show how this body of theory can be understood as generalizing the Bayesian approach to prediction.

The second step (sect. 3.2) is the empirical success of object-induction: here I will discuss the problem of the description of induction, and argue for a high-level view where the object-inductive strategy is identified with scientific method, that stands in competition with various nonscientific methods.

The conclusion of Schurz's argument (sect. 3.3) is that the optimal meta-inductive strategy confers justification to the most successful strategy so far, the object-inductive strategy. Here I will, first, discuss whether the meta-inductivist's optimality actually amounts to a *justification* for it. Second, I will discuss whether justification for the optimal meta-inductive strategy indeed goes over to justification for object-induction. I conclude in section 4.

2. The framework of prediction

In this section I set out and motivate the framework of sequential prediction that is presupposed throughout the paper. My presentation of this framework is fully based on existent theory, as reported in the textbook by Cesa-Bianchi and Lugosi (2006) and in Schurz's papers. I stay as close as possible to Schurz's presentation and notation, but also deviate from it where I think this makes things more clear.

2.1. The prediction game: events and predictions.

Schurz defines a prediction game as a pair $(\boldsymbol{y}^\omega, \Pi)$ of a *history* $\boldsymbol{y}^\omega$ and a pool Π of *prediction strategies*. It makes sense to extend this definition to include a third component, the *loss function*; this I will discuss in sect. 2.2.

A *history* $\boldsymbol{y}^\omega$ is an infinite sequence of *events*. Events are identified with values in some set Val of possible values; I will provide examples below. Write y_n for the n-th element of $\boldsymbol{y}^\omega$, or the event in *round* n of the game ($n \in \mathbb{N}^{>0}$).

Predictions are elements in some set Val_{pred}; again I will provide examples below. A *prediction strategy*, an element of the pool Π, specifies in each round a prediction about the next event. Write $\text{pred}_n(P)$ for the prediction of strategy P for round n.

Schurz in (2008; 2009) discusses two different types of prediction games: the *binary* and the *real-valued* prediction game.

2.1.1. *Example: the binary prediction game.*

Here the events can take one of two possible values, 0 and 1; and the predictions are simply categorical guesses about the possible values, 0 or 1. That is, $\text{Val} = \text{Val}_{\text{pred}} = \{0, 1\}$.

This is the prediction game Schurz (2004) discusses in an early exposition of his proposal. However, this simple prediction game is complicated in the sense that it is impossible to specify optimal strategies, as explained below. This is perhaps why Schurz in (2008; 2009) also considers the following more general game.

2.1.2. *Example: the real-valued prediction game.*

In this game both the events and the predictions take real values in the range $[0, 1]$. That is, $\text{Val} = \text{Val}_{\text{pred}} = [0, 1]$.

As explained in more detail below, for this game we can establish optimality results. The reason is that we can get the most out of *hedging* strategies, that predict by taking some weighted mean of the predictions of all the strategies in the pool: namely, for real-valued predictions, and unlike binary predictions, any weighted mean is still a valid prediction.

2.1.3. *Example: the probabilistic binary prediction game.*

But this we can already achieve by allowing real-valued predictions in the binary prediction game. Formally this gives the game where $\text{Val} = \{0, 1\}$ and $\text{Val}_{\text{pred}} = [0, 1]$.

A likely interpretation is that we have introduced the possibility of *randomization* in the binary prediction game: a prediction strategy now has access to a random device, which in each round it can use to toss a virtual coin with any desired bias $p \in [0, 1]$, and give out prediction 1 if it lands heads and 0 if it lands

tails. Note, however, that on this interpretation we are still in the categorical binary prediction game; the difference is that we have allowed prediction strategies to be random variables. Concretely, this means that the bounds we derive below are applicable to the binary prediction game only *in expectation* (see Cesa-Bianchi and Lugosi, 2006, 67ff).

The alternative interpretation is that a strategy directly issues *probabilistic* rather than categorical predictions. That is, a prediction strategy issues elements in $\mathrm{Val}_{\mathrm{pred}} = [0, 1]$, where we interpret the latter as probabilities (for the next outcome being 1, say). In particular, the *loss function* scores the difference between the actual outcome and the issued probability, sect. 2.2 below.

This *probabilistic* binary prediction game constitutes, I think, a natural general framework of prediction (also see sect. 2.4 below). It is therefore noteworthy that this game does not receive much attention in Schurz's published papers: less attention, in any case, than possible ways of simulating the hedging strategy in the game where predictions *must* be categorical. Thus Schurz (2017, 828f; 2018, 3890) considers two possibilities: the aforementioned introduction of randomnization (2016, 48), and the use of a *collective* of meta-inductivists, the categorical predictions of which we then take the mean value of (2008, 298ff; 2009, 214f). Schurz prefers the second solution to the interpretation of randomnization, making the valid point that the latter "is not entirely general since it presupposes that the events are probabilistically independent from [the meta-inductivist's] choice of prediction," (2017, 828; cf. Shalev-Shwartz and Ben-David, 2014, 252). Still, the introduction of a collective of meta-inductivists is ultimately a way of simulating single probabilistic predictions, and, due to rounding-off effects (see Schurz, 2008, 299; 2009, 215), an imperfect way at that. It seems to me that both the choice of game and Schurz's choice of evaluating the *mean* success of a collective of strategies are modeling choices (cf. Schurz, 2009, 215), that are less general and less elegant than evaluating the loss of a single meta-inductivist that is allowed probabilistic predictions.

2.2. **The prediction game: the loss function.** Strategy P, when making prediction $\mathrm{pred}_n(P) = \mathrm{pred} \in \mathrm{Val}_{\mathrm{pred}}$ for round n, suffers, when the outcome is revealed to be $y_n \in \mathrm{Val}$, a *loss* $\ell(\mathrm{pred}, y_n)$.

That is, a loss function $\ell : \mathrm{Val}_{\mathrm{pred}} \times \mathrm{Val} \to [0, \infty)$ quantifies how much a prediction was off in light of the actual outcome. Assuming that the function is nonnegative, that it returns 0 in case of the best possible prediction given the outcome, and that it is increasing in divergence between the prediction and the outcome, we are still left with much freedom in choosing a particular function.

2.2.1. *Example: the zero-one loss function.* Defined by $\ell_{0/1}(\mathrm{pred}, y) = 0$ if $\mathrm{pred} = y$ and $\ell(\mathrm{pred}, y) = 1$ otherwise, this is an obvious loss function in the categorical binary prediction game.

2.2.2. *Example: the absolute loss function.* Defined by $\ell_{\mathrm{abs}}(\mathrm{pred}, y) = |\mathrm{pred} - y|$, and called the *natural* loss function by Schurz, this is indeed an obvious loss function in the real-valued prediction game. Note, moreover, that this function reduces to the zero-one loss function in the categorical binary prediction game.

The absolute loss function can, of course, also be used in the probabilistic binary prediction game. In fact, in the categorical prediction game that allows strategies the use of randomnization, the *p-expected* zero-one loss coincides with the absolute loss $\ell_{\mathrm{abs}}(p, y) = |p - y|$ evaluated on probabilistic prediction $p \in [0, 1]$. It is in

this sense that one might say that the categorical binary prediction game with randomnization and the probabilistic binary prediction game coincide. However, to repeat, the difference is that in the first case we are dealing with *expected* losses, whereas in the second case the loss function deals with probabilistic predictions directly.

The absolute loss function does fail to satisfy a usual requirement on loss functions in a probabilistic context. Namely, the absolute loss function is not *proper*: whenever $p > 0.5$ (respectively, $p < 0.5$), the p-expected loss of prediction q is minimized by the extreme prediction $q = 1$ (by $q = 0$), rather than by p itself.

2.2.3. *Example: the square loss function.* Defined by $\ell_{\text{sq}}(\text{pred}, y) = (\text{pred} - y)^2$, this loss function *is* proper.

2.2.4. *Convex bounded loss functions.* The absolute and the square loss function are both *convex* in their first argument: for all $\gamma \in [0, 1]$, predictions $\text{pred}_1, \text{pred}_2 \in [0, 1]$ and outcomes y, it holds that

$$\ell(\gamma \cdot \text{pred}_1 + (1 - \gamma) \cdot \text{pred}_2, y) \leq \gamma \cdot \ell(\text{pred}_1, y) + (1 - \gamma) \cdot \ell(\text{pred}_2, y).$$

In words, the loss of a prediction that is a weighted mean of several predictions, is no worse than the weighted mean of the losses of these individual predictions. This is an important technical property that underlies the below results on the existence of optimal strategies; strategies that indeed proceed by taking weighted averages of all available predictions.

The absolute and the square loss function are also *bounded*, meaning that their range is the interval $[0, 1]$. The first main optimality result that I discuss below (sects. 3.1.2–3.1.4) covers all convex bounded loss functions.

2.2.5. *Example: the logarithmic loss function.* Defined by

$$\ell_{\log}(\text{pred}, y) = \begin{cases} -\ln(1 - \text{pred}) & \text{if } y = 0 \\ -\ln \text{pred} & \text{if } y = 1 \end{cases},$$

this is a prominent loss function (Cesa-Bianchi and Lugosi, 2006, 247ff), for its technical properties as well as its information-theoretic interpretation (see, e.g., Grünwald, 2007; Sterkenburg, 2018, 137ff), and, as I will discuss in sect. 3.1.6 below, for its strong connection to *Bayesian* prediction.

The logarithmic loss function is proper and convex, but it is *not* bounded. (Indeed, the loss of extreme prediction $\text{pred} = 0$ on outcome $y = 1$, and $\text{pred} = 0$ on $y = 1$, is *infinity*.) As such, it falls outside of the scope of the below optimality result on convex bounded loss functions—but this is no matter, because, as I will explain below too (sects. 3.1.5–3.1.7), it actually enables *stronger* results.

2.2.6. *Cumulative losses and loss rates.* Schurz defines the *score* of strategy P in round n by $s(\text{pred}_n(P), y_n) := 1 - \ell(\text{pred}_n(P), y_n)$, and the *absolute success* $\text{abs}_n(P)$ of P by the conclusion of round n as the sum $\sum_{i=1}^{n} s(\text{pred}_i(P), y_i)$ of its scores up to n. It is, however, more usual and indeed more convenient to evaluate things directly in losses, so that I will rather consider the absolute or *cumulative* loss given by $\text{Loss}_n(P) := \sum_{i=1}^{n} \ell(\text{pred}_i(P), y_i)$. Likewise, I define (parallel to Schurz's *success rate*) the *loss rate* $\text{loss}_n(P)$ of P by n as the average $\text{Loss}_n(P)/n$ of its losses up to n.

2.3. **The goal: an optimal strategy.** Given a pool Π of prediction strategies, we aim to design a *meta-inductive* strategy MI that, having access to the predictions of all the other strategies, predicts in such a way that it is *optimal* with respect to Π. To a first approximation, strategy MI is optimal with respect to Π if by following it we will *always* do about as good as we possibly could have done, given that the strategies in Π represent what we could have done. In other words, MI is optimal if it will *always* be about as successful as the most successful strategy in the pool. Here 'always' means: on *every* history of events. Importantly, we make *no* assumption whatsoever on possible histories. Another way of saying this is that we aim to derive *worst-case* guarantees.

What should 'about as successful' mean? As a start, we can demand that MI is at least as successful as any $P \in \Pi$ *in the long run*. Let us define (parallel to Schurz's maximal success rate) by $\text{minloss}_n := \min_{P \in \Pi \cup \{\text{MI}\}} \text{loss}_n(P)$ the minimum loss rate among all the strategies (including the meta-inductivist itself) by round n. Then we can demand that MI's loss rate—its average loss per round—converges to no more than the lowest loss rate achieved by any of the strategies, in the long run:

$$\lim_{n \to \infty} (\text{loss}_n(\text{MI}) - \text{minloss}_n) = 0. \tag{1}$$

(Again, there is no mention of the actual history: this is because it should hold for *all* histories.)

We will see below that we can indeed design strategies that achieve (1), at least for finite Π. This is still, however, a relatively weak demand: long-run convergence does not guarantee anything about the performance of the meta-inductivist by any given finite time. It does not guarantee anything about performance in the short run—or the *speed* of convergence.

We can express speed of convergence in terms of a function f, that depends on n and inevitably also on the size $K := |\Pi|$ of the pool of strategies, such that for all rounds n,

$$\text{loss}_n(\text{MI}) \leq \text{minloss}_n + f(n, K). \tag{2}$$

The function f bounds what we call the *regret rate*: it bounds MI's *surplus* loss rate up to n, as compared to the loss rate it *would* have incurred if it would always have mimicked the strategy in Π that was, in hindsight, the most successful up to n. The regret rate, to give an interesting optimality bound, needs to be decreasing in n (or it would not even guarantee (1))—and the faster the better.

Work in the field of prediction with expert advice has indeed led to meta-inductive algorithms that satisfy (2) for quickly decreasing f (Cesa-Bianchi and Lugosi, 2006). For the general class of convex bounded loss functions we can achieve an $f(n, K)$ with a main term of the form $\sqrt{\ln K / n}$, giving a regret rate that decreases in $1/\sqrt{n}$ (sects. 3.1.2–3.1.4). Moreover, for an important subclass of convex loss functions, f is of the form $(\ln K)/n$, giving a regret rate that decreases in $1/n$ (sects. 3.1.5–3.1.7). The latter indeed entails a finite constant bound (only depending on K) on the difference between MI's and any P's *total* losses, or, for any P, the *cumulative regret* relative to P (of MI for not having mimicked P from the start)

$$\text{Regret}_n(P) := \text{Loss}_n(\text{MI}) - \text{Loss}_n(P). \tag{3}$$

In all of the following, when I loosely use the term 'optimality,' it is these kind of short-run guarantees that I have in mind.

2.4. **The prediction game and the problem of induction.** The previous set out the framework of the prediction game. It is within the confines of this framework that Schurz's proposed justification of induction proceeds. But, of course, there is the concern that this framework of prediction falls short of capturing the essence of inductive inference, and so a justification within this framework falls short of the real goal: to justify inductive or scientific reasoning (cf. Arnold, 2010, 591).

One can, to begin with, object that scientific inquiry consists not so much in producing forecasts as in inferring general conclusions: not so much in prediction as in the formulation and the confirmation of hypotheses and theories. One can thus object to the generality of the problem setting of prediction; one can further object to the way predictive inference is rendered in our framework. Dawid (1984, 279), when introducing what is essentially our probabilistic prediction game under the header of "prequential forecasting," is frank about its limitations: "[t]he data may arrive *en bloc*, rather than in a natural order; if they come from a time-series, it may be impossible, or not obviously desirable, to analyse them at every point of time, or to formulate one-step ahead forecasts; and the restriction whereby all uncertainty about the next observation is to be encoded in a probability distribution, while acceptable to Bayesians, may not appeal to others."

Nevertheless, while there certainly is a case to be made that prediction is only a subsidiary part of science, there is also an important opposed tradition, indeed fashionable in many parts of machine learning today, that takes it that scientific inference ultimately comes down to inductive inference from particulars to particulars, or predictive inference. Moreover, while our framework certainly cannot accommodate everything there is to say about prediction, and particularly the choice of a loss function introduces an arguably highly context-dependent element, I think, and will assume in this paper, that the probabilistic prediction game, in which we can demonstrate the desired optimality for a wide class of loss functions, possesses a level of generality that lends significance to the conclusions we draw from it: ultimately, hopefully, an actual justification of induction.

3. The argument

The proposed justification of induction proceeds by two steps: an analytical proof that meta-induction, which favors the most successful available strategy, is an optimal strategy (sect. 3.1), and the empirical fact that object-induction has so far been the most successful strategy (sect. 3.2). From this is to follow the conclusion that the optimal strategy would now proceed object-inductively, thus justifying object-induction (sect. 3.3).

3.1. **Step one: the analytical optimality of meta-induction.** Here I explain the "mathematical-analytic" justification of meta-induction (Schurz, 2008, 282), that amounts to the definition of a meta-inductive prediction strategy and a theorem about its performance. Again, my presentation of the relevant mathematics is fully based on existing results in the field of prediction with expert advice.

3.1.1. *Towards optimality: follow the leader or hedge your bets?* The most straightforward meta-inductive strategy is the strategy that in each round simply issues

the same prediction as the then most successful strategy in the pool (Schurz's "simple meta-inductivist," 2008, 285ff; or "imitate-the-best," 2009, 207; elsewhere also "follow-the-best-expert," Cesa-Bianchi and Lugosi, 2006, 41ff; or simply "follow-the-leader," De Rooij et al., 2014, 1282). This simple strategy often works well; specifically, it approaches the leader's success if there are not too many alternations in *which* strategy is the leader (ibid., 1295; generalizing Schurz's observation, 2008, thm. 1; 2009, thm. 1, that it works well if from some round on one and the same strategy will always be best). However, in the *worst case* this strategy can be very bad. It is very bad in case of an adversarial history that forces an endless rapid change in which strategy is the best, so that the meta-inductivist is always switching to a strategy that will *then* issue a bad prediction, making it perform significantly worse than each strategy in the pool (Schurz, 2008, 285ff; Schurz, 2009, 208; De Rooij et al., 2014, 1282).

In fact, a similar kind of scenario shows that in the *categorical* binary prediction game, *no* meta-inductive strategy can be optimal. Namely, in the categorical game, recall, strategies must give as predictions either 0 or 1; so (assuming that in each round there are at least two strategies with different predictions) the meta-inductivist must each round wholeheartedly follow the predictions of some strategy or strategies, and an adversarial history can then make *these* strategies fail, while the others in the pool are in that round maximally successful. So already in the simplest scenario of the two constant strategies that always make the same prediction, an adversarial history can make the meta-inductivist fail *in each round*, whichever strategy it then follows, while clearly for each possible history at least one of the strategies must at least half of the time predict correctly (Cesa-Bianchi and Lugosi, 2006, 67; Schurz, 2008, 298f.)

The method to avert such scenarios—a method unavailable in the categorical game, but unlocked in the probabilistic game—is to 'hedge your bets.' That is, you predict in each round by some weighted average of the predictions of all the elements in the pool. This gives the *weighted-average meta-inductivist* wMI defined by

$$\text{pred}_{n+1}(\text{wMI}) := \sum_{P\in\Pi} w_n(P)\cdot \text{pred}_{n+1}(P). \tag{4}$$

From this perspective (the "hedge setting," De Rooij et al., 2014, 1282, in terminology going back to Freund and Schapire, 1997), the challenge comes down to choosing appropriate weights. It turns out that, with the right choice of weights, that depend on the strategies' losses, any adversarial history that inflicts a high loss rate on the meta-inductivist *must also inflict a high loss rate on all the other strategies*. That is, such a weighted-average meta-inductivist always keeps its *regret rate* low: it is an optimal strategy.

3.1.2. *The attractiveness-weighted meta-inductivist.* The weighted-average strategy first discussed by Schurz (2008, 296; 2009, 213) is the *attractiveness-weighted meta-inductivist*, denoted awMI. The weight it assigns to strategy P after round n is determined by how much better P did in hindsight: what Schurz calls P's *attractiveness* but what is better known as the meta-inductivist's cumulative regret $\text{Regret}_n(P)$ relative to P, defined by (3) in sect. 2.3 above. The attractiveness-weighted meta-inductivist only takes into account those P that actually were more

successful than it; those that were not, hence would give regrets that are nonpositive, we can conveniently exclude with the help of the abbreviation $(\text{Regret}_n(P))_+ := \max\{0, \text{Regret}_n(P)\}$. The predictions of awMI are then defined, for $n \in \mathbb{N}$, by

$$\text{pred}_{n+1}(\text{awMI}) := \frac{\sum_{P\in\Pi} (\text{Regret}_n(P))_+ \cdot \text{pred}_{n+1}(P)}{\sum_{P\in\Pi} (\text{Regret}_n(P))_+}. \tag{5}$$

(Where we can stipulate $\text{Regret}_0(P) = 1$, say, for all P, so that in the first round these terms simply cancel out.) This is the weighted-average meta-inductivist (4) with weights

$$w_n(P) = \frac{(\text{Regret}_n(P))_+}{Z}, \tag{6}$$

where Z always denotes a normalization term, in this case $Z = \sum_{P\in\Pi} (\text{Regret}_n(P))_+$.

One can show, for all bounded convex loss functions, that awMI satisfies the long-run convergence bound (1). Namely, one can show (Cesa-Bianchi and Lugosi, 2006, cor. 2.1, Schurz, 2008, thm. 4; 2009, thm. 2) that awMI satisfies the bound (2) for f of order $1/\sqrt{n}$; specifically, for all bounded convex loss functions,

$$\text{loss}_n(\text{awMI}) \leq \text{minloss}_n + \sqrt{\frac{K}{n}}. \tag{7}$$

3.1.3. *Improvements: polynomial weights.* As noted by Schurz (2008, 297), the attractiveness weighted strategy is a special case of the *polynomially*-weighted-average strategy in Cesa-Bianchi and Lugosi (2006, 12), where the unnormalized weights are given by

$$(\text{Regret}_n(P))_+^\eta \tag{8}$$

for some parameter $\eta \geq 1$. Indeed, the bound (7), hence the worst-case speed of convergence, can be improved by a different choice of parameter. In the words of Cesa-Bianchi and Lugosi (2006, 13, notation mine), "The choice $\eta = 1$ yields a particularly simple algorithm. On the other hand, the choice $\eta = 2\ln K - 1$ (for $K > 2$) ... leads to $\text{loss}_n(\text{awMI}) \leq \text{minloss}_n + \sqrt{e(2\ln K - 1)/n}$, yielding a significantly better dependence on the number of experts K." Namely, the factor in the bound introduced by the size of the pool is now only its logarithm.

3.1.4. *Improvements: exponential weights.* The central strategy in the field, however, is the *exponentially*-weighted-average strategy, that employs unnormalized weights of the form

$$\exp(\eta \cdot \text{Regret}_n(P)) \tag{9}$$

for parameter $\eta > 0$ (the *weighted-majority algorithm*, Littlestone and Warmuth, 1994; see Cesa-Bianchi and Lugosi, 2006, 14f; Shalev-Shwartz and Ben-David, 2014, 252ff; and Schurz, 2009, 214; Schurz and Thorn, 2016, 52). Note that (9) equals

$$\exp(\eta \cdot (\text{Loss}_n(\text{wMI}) - \text{Loss}_n(P))) = \exp(\eta \cdot \text{Loss}_n(\text{wMI})) \cdot \exp(-\eta \cdot \text{Loss}_n(P)),$$

and the first factor $\exp(\eta \cdot \text{Loss}_n(\text{wMI}))$ cancels out in the normalization. Consequently, the choice of unnormalized weights (9) is equivalent to the choice

$$\exp(-\eta \cdot \text{Loss}_n(P)). \tag{10}$$

The exponentially-weighted strategy thus has the special property that its weights and hence its predictions *only* depend on the other strategies' past successes, and not also on its *own* past success (Cesa-Bianchi and Lugosi, 2006, 14).

Parameter η is also called the *learning rate*: the larger it is, the greater the inflation of differences in losses, and the more aggressive the updating. In this case the optimal way of tuning the parameter η is to let it decrease in a specific way in the number of rounds (ibid., 17f). This still gives a bound with a main term of the form $\sqrt{\ln K/n}$, which is indeed provably the best we can hope to achieve in general for the convex bounded loss functions (see ibid., 62f).

3.1.5. *Improvements for other loss functions.* The *very* strongest bound we could ever hope to derive, *whatever* the loss function, is of the form $(c \ln K)/n$ for some constant c—that is, a bound of order $1/n$. Namely, for any given loss function, one can design a pool of experts and a history such that any meta-inductive strategy's loss rate falls at least this much short of the best strategy's (see ibid., 59ff).

As a matter of fact, we *can* derive this guarantee for a particular class of loss functions, that includes, among others, the quadratic and the logarithmic loss. I will now first discuss the important particular case of optimal probabilistic prediction with the logarithmic loss function, which is all the more interesting for its natural Bayesian interpretation.

3.1.6. *The Bayesian meta-inductivist and the logarithmic loss function.* In the probabilistic binary prediction game, recall from sect. 2.1.3, strategies in each round issue a probability $\text{pred} \in [0,1]$ of the next outcome being 1. As such, they can be seen to give conditional *likelihoods* (where $n \in \mathbb{N}$, $\boldsymbol{y}^n$ denotes the event sequence up to round n, and by stipulation $P(\cdot \mid \boldsymbol{y}^0) := P(\cdot)$)

$$P(y_{n+1} \mid \boldsymbol{y}^n) = \begin{cases} 1 - \text{pred}_{n+1}(P) & \text{if } y_{n+1} = 0 \\ \text{pred}_{n+1}(P) & \text{if } y_{n+1} = 1 \end{cases}, \tag{11}$$

and so, by the chain rule of conditional probabilities

$$P(\boldsymbol{y}^n) := \prod_{i=0}^{n-1} P(y_{i+1} \mid \boldsymbol{y}^i), \tag{12}$$

likelihoods over finite histories (Cesa-Bianchi and Lugosi, 2006, 247f).

The *Bayesian* meta-inductivist P_{BayMI} maintains a probability model over both the space of possible histories and the pool of strategies. First, the marginal prior probability of each of the strategies is stipulated by some initial weight function w_0; thus $P_{\text{BayMI}}(P) := w_0(P)$. Second, the strategies' likelihoods are stipulated to determine conditional probabilities $P_{\text{BayMI}}(\boldsymbol{y}^n \mid P) := P(\boldsymbol{y}^n)$ of histories. This, then, suffices to determine the marginal prior probability of histories. Namely, via the law of total probability we have

$$P_{\text{BayMI}}(\boldsymbol{y}^n) = \sum_{P \in \Pi} w_0(P) \cdot P(\boldsymbol{y}^n). \tag{13}$$

Now the Bayesian meta-inductivist's predictions for round $n+1$ are, by Bayes's rule, simply given by the conditional probabilities on the past events by n,

$$P_{\text{BayMI}}(y_{n+1} \mid \boldsymbol{y}^n) = \frac{P_{\text{BayMI}}(\boldsymbol{y}^{n+1})}{P_{\text{BayMI}}(\boldsymbol{y}^n)}. \tag{14}$$

Alternatively but equivalently (cf. Grünwald, 2007, 77), for each round $n+1$ we can first calculate the *posterior* marginal probabilities by this round, identified

with the weights w_n by this round, that the Bayesian meta-inductivist assigns to the strategies. By Bayes's rule, we update the new weights after round n to

$$w_n(P) := w_0(P \mid \boldsymbol{y}^n) = \frac{w_0(P) \cdot P(\boldsymbol{y}^n)}{Z}, \tag{15}$$

with normalization $Z = \sum_{P \in \Pi} w_0(P) \cdot P(\boldsymbol{y}^n)$. Then by the law of total probability

$$P_{\text{BayMI}}(\cdot \mid \boldsymbol{y}^n) = \sum_{P \in \Pi} w_n(P) \cdot P(\cdot \mid \boldsymbol{y}^n). \tag{16}$$

In particular,

$$\text{pred}_{n+1}(\text{BayMI}) = \sum_{P \in \Pi} w_n(P) \cdot \text{pred}_{n+1}(P). \tag{17}$$

In short, the Bayesian meta-inductivist for the probabilistic prediction game is the weighted-average meta-inductivist with weights given by Bayes's rule, (15).

An important difference, at first sight, from the attractiveness-weighted meta-inductivist awMI, is that the weights of BayMI depend on the strategy's probabilistic *predictions* and so only indirectly on their *losses*. However, the Bayesian meta-inductivist is strongly connected to the logarithmic loss function (Cesa-Bianchi and Lugosi, 2006, 247ff), which can be seen by writing out

$$\begin{aligned}
\text{Loss}_n(P) &= \sum_{i=0}^{n-1} \ell(\text{pred}_{i+1}(P), y_{i+1}) \\
&= \sum_{i=0}^{n-1} -\ln P(y_{i+1} \mid \boldsymbol{y}^i) \\
&= -\ln \prod_{i=0}^{n-1} P(y_{i+1} \mid \boldsymbol{y}^i) \\
&= -\ln P(\boldsymbol{y}^n).
\end{aligned}$$

That is, for the logarithmic loss, $P(\boldsymbol{y}^n) = \exp(-\text{Loss}_n(P))$. Substituting this term in the weight specification (15), we obtain

$$w_n(P) = \frac{w_0(P) \cdot \exp(-\text{Loss}_n(P))}{Z}. \tag{18}$$

Another apparent difference with wMI is that BayMI still explicitly depends on the *prior* weights given by w_0. However, since we assume a finite pool, we can choose *uniform* prior weights $w_0(P) = 1/K$, in which case they all cancel out and we finally obtain

$$w_n(P) = \frac{\exp(-\text{Loss}_n(P))}{Z}. \tag{19}$$

In words, the (uniform-prior) Bayesian meta-inductivist for the logarithmic loss function is the exponentially-weighted meta-inductivist introduced in sec. 3.1.4 above, with parameter $\eta = 1$ (ibid., 250).

Now it is easy to derive a bound on $\text{Loss}_n(\text{BayMI}) = -\ln P_{\text{BayMI}}(\boldsymbol{y}^n)$, using the simple fact that $\sum_{P \in \Pi} w_0(P) \cdot P(\boldsymbol{y}^n) \geq w_0(P) \cdot P(\boldsymbol{y}^n)$ for any particular strategy

P (ibid.). Namely,

$$\begin{aligned}\text{Loss}_n(\text{BayMI}) &= -\ln \sum_{P\in\Pi} w_0(P)\cdot P(\boldsymbol{y}^n)\\ &\leq -\ln\left(w_0(P)\cdot P(\boldsymbol{y}^n)\right)\\ &= -\ln w_0(P) + \text{Loss}_n(P).\end{aligned}$$

In particular, with uniform choice of initial weights, we have for the loss rate that

$$\text{loss}_n(\text{BayMI}) \leq \text{minloss}_n + \frac{\ln K}{n}. \tag{20}$$

3.1.7. *Mimicking the Bayesian meta-inductivist for other loss functions.* In the same way that we evaluated the cumulative loss

$$\text{Loss}_n(\text{BayMI}) = -\ln \sum_{P\in\Pi} w_0(P)\cdot \exp(-\text{Loss}_n(P)) \tag{21}$$

in the special case of the Bayesian meta-inductivist with the *logarithmic* loss (the exponentially-weighted strategy with $\eta = 1$), we can, for *any* given loss function ℓ, as well as any learning rate $\eta > 0$, evaluate the term

$$\text{MixLoss}_n := -1/\eta \cdot \ln \sum_{P\in\Pi} w_0(P)\cdot \exp(-\eta\cdot \text{Loss}_n(P)), \tag{22}$$

the so-called cumulative *mix-loss*. In particular, much like before, we can derive for *any* loss function that

$$\text{MixLoss}_n \leq -1/\eta \cdot \ln w_0(P) + \text{Loss}_n(P), \tag{23}$$

so that, for uniform weights, the mix-loss *rate* satisfies

$$\text{mixloss}_n \leq \text{minloss}_n + \frac{\ln K}{\eta\cdot n}. \tag{24}$$

But what does this actually mean? Well, for some loss functions ℓ, the term (22), the mix-loss, is strictly meaningless, which is to say that there is no actual (meta-inductive) strategy that this cumulative loss term corresponds to. But for an important subclass of convex loss functions, the so-called *mixable* loss functions, and the right choice of η, the mix-loss (22) corresponds to an actual strategy (the *aggregating strategy* due to Vovk, 1990, also see Vovk, 1998, 2001; Cesa-Bianchi and Lugosi, 2006, 52ff; Grünwald, 2007, 573ff; Sterkenburg, 2018, 143ff). This meta-inductive strategy aMI can be seen as explicitly mimicking the Bayesian strategy with the logarithmic loss function, for *this* loss function, with the aim of achieving the Bayesian performance bound (20). And indeed, virtually by definition, aMI is optimal for the given mixable loss function in the strong sense of (24).

The mixable loss functions include such important functions as the logarithmic loss (obviously: the mix-loss with $\eta = 1$ is the Bayesian strategy) and the square loss, but not the absolute loss function.

3.1.8. *Summary.* We have considered a number of central meta-inductive strategies for general classes of loss functions. These meta-inductive strategies are optimal for the relevant prediction games; they not only satisfy long-run convergence to the minimal loss rate among all strategies in the pool, but indeed short-run bounds of order $1/\sqrt{n}$ (for the convex bounded loss functions) or even of order $1/n$ (the mixable loss functions).

We also saw that all of these optimal meta-inductive strategies are weighted-average strategies (or, for some mixable games, aggregating strategies), that predict by combining weighted predictions of all the other strategies in the pool. We indeed saw that the exponentially-weighted strategies (optimal for the convex bounded loss functions) and the aggregating strategies (for the mixable loss functions) can be understood as generalizations of the standard Bayesian strategy for the specific case of the logarithmic loss function.

Moreover, the assigned weights in all cases depend on the strategies' losses, that is, on their past performance. Now in the Bayesian case the weights also depend on a *prior* weight assignment w_0, and we can introduce such a prior weight factor in the other meta-inductive strategies, too (see, e.g., Cesa-Bianchi and Lugosi, 2006, exc. 2.5): but the effect of these weights cancels out in case of a *uniform* prior over the finite pool of strategies, and I will assume this to be the reasonable choice (see sect. 3.3.1 below). Under that stipulation, the weights depend on the strategies' performance *only*, and strategies that have been more successful receive a higher weight and have a larger share in the meta-inductivist's prediction. It is in that sense that we can say that the meta-inductive strategy *favors* strategies to that extent in which they have been more or less successful than the other strategies so far.

Thus, abstracting away from which of the standard loss functions we are actually assuming, and hence which precisely is the strategy that we will now simply call 'the meta-inductive strategy MI,' we can formulate the first, analytical step of Schurz's argument as

(A) The meta-inductive strategy MI, that at each point in time favors strategies to the extent of their relative success so far, is an optimal strategy.

3.2. **Step two: the empirical success of object-induction.** The second step is the empirical observation that "so far object-induction has turned out to be the most successful prediction strategy" (Schurz, 2008, 304).

3.2.1. *The object-inductive strategy?* The problem of the justification of induction looks much intertwined with that of the *description* of inductive method (see Lipton, 2004, 7ff; Skyrms, 2000, 51ff; Goodman, 1954, 68ff). Thus Schurz in (2004) starts by asking "*which* inductive method should be attempted to justify," and then notes that "there is a fundamental method that more or less underlies all particular inductive methods – the *straight rule*" (243, translation mine). In subsequent papers (2008, 283f; 300f; 2009, 206), Schurz also discusses the straight rule and refined version thereof.

Even if Schurz's purpose here is merely to illustrate some basic inductive strategies and how to refine them in more complex environments, it does bring out the worry that the proposed meta-inductive justification is conditional on some satisfying solution to the problem of description. But this is itself a fundamental problem; certainly any conception of object-induction as some variant of the straight rule directly faces Goodman's observation that induction is underdetermined in which past regularities are actually extrapolated. Of course, we have assumed a set of possible events, and the straight rule extrapolates the past relative frequencies of *these* events; but that merely shifts the problem to whether and how we have managed to carve out possible events in the world that are indeed extrapolatable or *projectible* in this way. Schurz (2008, 279) writes, "I will avoid Goodman's 'new

riddle' of induction by assuming that the events are explicated in terms of qualitative predicates or function symbols in the sense of Carnap (1947, 146)." But this is a weak response: Carnap's reply was found wanting from the start (Goodman, 1947; 1954, 78ff; Putnam, 1983, xf).

3.2.2. *Meta-induction and entrenchment.* So how to deal with the problem of description in the context of Schurz's proposed justification? I can see two possible ways out.

One is to face the problem of description head-on. Perhaps the optimality of meta-induction can be employed to found an account of projectibility of regularities in terms of their past successful projections or their *entrenchment*, much in the spirit of Goodman's own suggestions (1954, 84ff). Thus the pool of prediction strategies is taken to represent the candidate projectable regularities, and we are meta-inductively justified to favor the strategy most successful so far, or the regularity best entrenched. Of course, this is all fairly speculative: much more work would need to be done to make this idea precise, and I will not attempt to do that in this paper.

3.2.3. *The scientific method.* Rather, my proposed way out goes to the other extreme: I will argue that in the context of Schurz's argument, we can simply *disregard* the problem of description.

What is crucial for the force of Hume's skeptical argument is that it goes through for any inductive method one could propose. The problem of induction can therefore actually be stated quite independently of the problem of description (Lipton, 2004, 11). But something similar holds for Schurz's constructive argument. The argument does not rely on any specific characteristics that the object-inductive method might or might not have—it only relies on *its past success*. For Schurz's proposed justification of induction it is enough to accept that there is something like an inductive method, and that it has been successful in the past.

In fact, it is fair to say that the problem of induction derives much of its relevance from the identification of induction with *scientific* method, which turns the problem of induction into the problem of providing rational grounds for scientific procedure. This would indeed "not only be of fundamental epistemological importance; it would also be of fundamental cultural importance as part of the enterprise of enhancing scientific rationality (Schurz, 2008, 280; see, especially, Salmon, 1967, 54ff on the significance of the problem of induction). Now most of us will agree that there is something like scientific method, despite the fact that it is notoriously hard to make perfectly precise what it consists in. There is such a thing—and it can be distinguished from several *non*scientific approaches (like consulting horoscopes, or accepting the forecasts of demagogues). Moreover, most of us will agree that science's past predictions have been quite successful—more successful at least than the past predictions of these alternative methods. Then the major challenge is to turn this fact of science's past success into a convincing argument for sticking to it, rather than turning to alternative methods.

And *this* is what Schurz's argument promises to do. Adopting a high-level perspective, where we take as object-inductive strategy OI the scientific method, which we further take to stand in competition with a pool Π of alternative nonscientific strategies, we plausibly have the empirical fact that OI has been most successful so far; and the promise of Schurz's argument is that this can then be plugged in with

the analytical optimality (A) to lead to the desired justification of the scientific method.[1]

3.2.4. *Summary.* In short, I take the second step to be that

(E) As a matter of empirical fact, the object-inductive strategy OI, that we identify with the scientific method (and that we take to be in competition with various proposed nonscientific methods), has been, at this point in time, the most successful prediction strategy (among the pool Π of all of these competing strategies).

3.3. **Conclusion: meta-induction favors object-induction.** From the analytical observation (A) and the empirical observation (E) it follows that

(C) The meta-inductive strategy MI for the pool Π of OI and its nonscientific competing strategies, an optimal strategy for Π, favors most, at this point in time, the object-inductive strategy OI.

Does conclusion (C) amount to the desired justification of OI? Only if we are clear about at least two additional things: (1) the optimality of MI in fact amounts to a justification for MI, and (2) the justification for MI is conferred to the strategy that it favors most, in particular, OI.

3.3.1. *The justification for an optimal strategy.* An initial worry is that a Reichenbachian optimality justification, in its original form often referred to as the *pragmatic* justification of induction, is only that: a pragmatic justification. In my view, however, this denotation is a misnomer, insofar it suggests that a 'pragmatic' justification is not a proper, *epistemic*, justification. While a justification from optimality falls short of a justification from reliability, *guaranteed success*, it is still an epistemic justification: *guaranteed success whenever success can be achieved at all*. (Cf. Schurz, 2008, 282.)

Of course, this presupposes that we use an epistemically relevant notion of success. In the previous we equated success with low losses, and then made the notion of optimality precise in term of long-term and short-term relative success, or long-term and short-term constraints on the regret.

Now success whenever success can be achieved at all *in the long run*, or asymptotic convergence to a regret rate of 0, already gives a nontrivial notion of optimality. It cannot be attained, over all possible pools, by any purely object-inductive strategy which predictions will not depend on the other strategies in the given pool. This is essentially what Schurz calls the *dominance* of meta-induction over object-induction (2017, 829f); and he writes that Reichenbach "has pointed out that already the *optimality* argument may be considered as a sufficiently strong justification of meta-induction, insofar as meta-induction is the *only* prediction strategy for which optimality can be *rationally demonstrated*" (2008, 303).

[1] One might object that it is more natural to view 'scientific method' as a *family* of different object-inductive strategies, rather than as a single strategy. This would indeed be closer to Schurz's own view, see sect. 3.2.1 above. I think this view would make the discussion in sect. 3.3 below somewhat more involved but not different in essence; for this reason I will stick in this paper to the view of a single OI.

Recall, though, that Reichenbach's own notion of long-run optimality included infinitely many different strategies, and this is also the case for our notion. Concretely, given an optimal strategy, we can devise infinitely many other strategies that up to any given round issue any given predictions, before settling on the same predictions as the original strategy. (Thus we can, for any given amount of cumulative loss, devise a strategy that initially incurs at least this amount of loss on some history, yet that is still optimal.) That means that if we seek justification for an optimal strategy's *actual predictions*, which will be relevant for the next step of deriving justification for following OI's predictions, sect. 3.3.2 below, then the long-run criterion is not enough. If a strategy's actual predictions were justified solely by virtue of the strategy's long-run success, then by the above we would have the perverse result that in each round any possible prediction is justified (cf. Salmon, 1957).

Fortunately, the *short-term* regret bounds for meta-inductive strategies exclude the possibility of arbitrary short-time regrets, hence predictions. To appreciate how restrictive they are (and the resulting notion of optimality is), let us have a closer look at the motivation for the nature of these bounds, particularly, their twofold *uniformity*.

First of all, the loss bound (2) is for the various meta-inductive strategies derived from the guarantee

$$\text{loss}_n(\text{wMI}) \leq \text{loss}_n(P) + f(n, K), \tag{25}$$

for all $P \in \Pi$, i.e., from a bound that is uniform over *all strategies*. This is also connected to the choice of uniform initial weights w_0 for the weighted-average (or aggregating) meta-inductivist. With *non*-uniform initial weights we would have long-run optimality as before, and we would even have short-term guarantees, but not of the uniform kind: in the short-term bound the term K would be replaced by a strategy-dependent term $1/w_0(P)$. With non-uniform weights, we must have assigned some strategies weights strictly smaller than $1/K$, which translates in a worse guarantee relative to these strategies, or a suboptimal performance if one of these strategies turns out to be the most successful. In short, a uniform bound presents the strongest possible *single* guarantee, *independent of* which strategy is most successful. And the choice of uniform weights is not just arguably the most reasonable choice in that the meta-inductivist should treat all competing strategies initially on a par, or should later evaluate strategies on their success only (sect. 3.1.8 above); it is the reasonable choice in that it comes with the strongest possible uniform short-term optimality guarantee.

Second of all, short-term optimality is a worst-case or uniform notion over *all possible histories*, as stipulated in sect. 2.3 above. But one can think of other ways of defining optimality in terms of loss bounds that are uniform over strategies yet *non*uniform over possible histories. Specifically, one can define a notion of optimality that allows a somewhat looser all-cases relative loss rate bound in return for a much stricter bound in specific cases of particularly regular or *easy* data. However, such a notion must be motivated by a ruling that some possible histories are somehow more likely or more relevant than others: as such, it constitutes an assumption that some possible uniformities hold a privileged status. But the elegance of Schurz's argument is precisely that it promises a justification of induction *without* any uniformity-of-nature assumptions. In particular, the first step, on the

optimality of meta-induction, should be free of such assumptions. And for this, worst-case optimality, a uniform bound over every possible history, appears the right notion.

Now, if we accept this requirement of uniform optimality, we can directly *define* the very best possible meta-inductive strategy, for a given loss function, as that strategy that (for a given pool of strategies) minimizes the worst-case regret for all strategies and all histories (the *minimax* regret, Cesa-Bianchi and Lugosi, 2006, 30ff). This theoretical strategy then gives the absolutely optimal predictions, for the given loss function; and if we take these predictions to be thereby fully *justified*, then results on various extents to which the absolutely optimal minimax regret coincides with the short-term optimality bounds we discussed (ibid., 59ff; 252ff) show that our meta-inductive strategy's actual predictions are to various extents justified, too.

I will not go further into the mathematical specifics, though, because the analytical details of the question of justification ultimately cannot (and need not) be separated from the *empirical* context of Schurz's argument. It is, for instance, sufficient for the argument to connect justification to a notion of optimality that is strong enough that all optimal strategies give, at this point in time, a sufficiently high weight to the object-inductive strategy, which is at least partly an empirical question: we will discuss this next, in sect. 3.3.2.

Furthermore, the notion of optimality is, of course, a relative notion, relative to a particular pool of strategies. That means that an optimal strategy can only be justified insofar the original choice of pool of strategies is appropriate, a proper rendition of all we could do. Now it seems there is an obvious such pool in the context of Schurz's argument, namely the pool of proposed alternative strategies that figures in the empirical step (E). But this may actually not settle it; I return to this important point in the conclusion of the paper, sect. 4.

3.3.2. *The justification for a strategy favored most by an optimal strategy.* Suppose that we are justified in predicting with the optimal strategy MI, at each point in time, and given, by (E), that optimal strategy MI, at this point in time, favors most the object-inductive strategy OI. Now it sounds plausible enough that we are justified in favoring OI, at this point in time. But this is not yet a justification for strategy OI.

To begin with, there is a difference between actually following and merely *favoring* a strategy, at any point in time. That a meta-inductivist assigns the highest weight to a strategy does not entail that their predictions coincide. For instance, the meta-inductivist might give a much higher weight to OI than to each of a large number K of competitors; but those competitors might collectively still take up most of the weight and issue a prediction opposed to the one issued by OI, so that the prediction issued by the meta-inductivist is also very different from OI's. In that case it sounds implausible that MI confers justification to following OI's prediction.

Note, however, that such cases are excluded with a sufficiently strong empirical step (E). They are excluded by an (E) that says that, as a matter of empirical fact, the strategy OI has been so much more successful than its competitors, that the meta-inductivist attributes it such a large share of the total weight that its prediction (approximately) coincides with OI's prediction. Indeed, in that clear-cut case all optimal strategies would (approximately) give this same prediction,

resolving remaining doubts about the justification for the specific predictions of either of them (sect. 3.3.1 above). Since following MI's prediction is then justified, following OI's prediction, being (approximately) the same prediction, is justified, too. Thus a sufficiently strong (E) resolves the problem; but, of course, it remains up for debate whether, as a matter of empirical fact, OI has really been *that* successful (cf. Schurz, 2018, 3891).

This leads us to the final observation that we would at best have a justification for following OI's *current prediction*, which is not the same as a justification for the *strategy* OI. Here it clearly shows in what way Schurz's argument for the justification of object-induction is more subtle than trying to establish that object-induction is itself optimal, the original idea of Reichenbach (cf. Schurz, 2008, 281). Strategy OI is, again, *not* itself optimal. In order to still conclude that we are justified, from this point on, in always following strategy OI, we would need to establish that the optimal strategy from this point on always favors OI. This would be the case if OI from this point on is always (much) more successful than its competitors: but observation (E), that OI has been the most successful so far, does not, of course, entail *that*. We can easily come up with possible futures in which OI loses its edge and some competitor strategy predicts vastly better (and so MI ceases to favor OI): this is just the failure of the naive meta-inductive justification of induction.

Thus Schurz's argument, if successful, only gives a justification for sticking to strategy OI *for now*: it only gives a justification for following its prediction *at this point in time*.[2] This is, to reiterate, strictly weaker than a *justification for object-induction*, which is normally understood as a justification for always following the strategy OI. Nevertheless, this would still be an important result. The relevance of the problem of induction resides for large part in the question whether we can give justification for following the scientific method *now*, in the face of several nonscientific alternatives. If the argument is indeed successful, it arguably brings us closer to a justification of induction than any argument before, and possibly the closest we can ever get.

4. Conclusion

In this paper, I reconstructed and evaluated Schurz's proposed meta-inductive justification of induction. I also took the opportunity to discuss some of the main observations from the field of prediction with expert advice, including the interpretation of the optimal meta-inductive algorithms as generalizations of Bayesian prediction strategies.

I argued that Schurz's argument should be seen as operating from a high-level perspective where we take it that there is something like the object-inductive or *scientific* prediction strategy, which, from its past success, we wish to justify as the preferred strategy among several proposed alternative strategies.

I pointed out the argument's need for a strong—perhaps *too* strong—empirical postulate that says that the scientific strategy has indeed been the most successful among these alternatives *by far*. Furthermore, I pointed out that the argument provides at best a justification for sticking with the scientific strategy *for now*, although I stressed that this would still constitute an important result.

[2]Or even extending to some rounds in the future, depending on how much of a lead it actually has on its competitors.

There is, however, one important issue that I did not fully address in the current paper. This is the choice of the pool of strategies that the meta-inductivist should actually be optimal *for* (sect. 3.3.1 above). Again, it would seem that there is an obvious pool in the context of Schurz's argument, namely the pool of the scientific strategy and its actually proposed alternatives. This necessarily being a *finite* pool of strategies, it would also render Schurz's argument immune against the charge of Arnold (2010) that it is not generally possible to have optimal strategies for *infinite* pools. However, as I will argue in a subsequent paper (Sterkenburg, 2019), it is not clear that the relevant optimality suffices for the *analytical* component (A) of the argument. Indeed, I will argue there that the analytical step requires a notion of optimality that is robust against an *expanding* pool of strategies, posing an important challenge still to Schurz's proposal.

References

E. Arnold. Can the best-alternative justification solve Hume's problem? On the limits of a promising approach. *Philosophy of Science*, 77(4):584–593, 2010.

R. Carnap. On the application of inductive logic. *Philosophy and Phenomenological Research*, 8(1):133–148, 1947.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

A. P. Dawid. Present position and potential developments: Some personal views. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, 147: 278–292, 1984.

S. de Rooij, T. van Erven, P. D. Grünwald, and W. M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014.

F. Dietrich and C. List. Probabilistic opinion pooling. In A. Hájek and C. Hitchcock, editors, *The Oxford Handbook of Probability and Philosophy*, pages 519–541. Oxford University Press, Oxford, 2016.

H. Feigl. De principiis non disputandum . . . ? On the meaning and the limits of justification. In M. Black, editor, *Philosophical Analysis*, pages 119–156. Cornell University Press, New York, NY, 1950.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

N. Goodman. On infirmities of confirmation-theory. *Philosophy and Phenomenological Research*, 8(1):149–151, 1947.

N. Goodman. *Fact, Fiction, and Forecast*. The Athlone Press, London, 1954.

P. D. Grünwald. *The Minimum Description Length Principle*. MIT Series in Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2007.

P. Herz. Kritische Bemerkungen zu Reichenbachs Behandlung des Humeschen Problems. *Erkenntnis*, 6:25–31, 1936.

P. Lipton. *Inference to the Best Explanation*. Routledge, London, second edition, 2004.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

H. Putnam. Foreword to the fourth edition of N. Goodman, *Fact, Fiction, and Forecast*, Harvard University Press, Cambridge, MA, 1983.

H. Reichenbach. Die logischen Grundlagen des Wahrscheinlichkeitsbegriffs. *Erkenntnis*, 3:401–425, 1933.

H. Reichenbach. *Wahrscheinlichkeitslehre: eine Untersuchung Über die Logischen und Mathematischen Grundlagen der Wahrscheinlichkeitsrechnung*. Sijthoff, Leiden, 1935.

H. Reichenbach. *Experience and Prediction*. University of Chicago Press, Chicago, IL, 1938.

W. C. Salmon. The predictive inference. *Philosophy of Science*, 24(2):180–190, 1957.

W. C. Salmon. *The Foundations of Scientific Inference*. University of Pittsburgh Press, Pittsburgh, PA, 1967.

G. Schurz. Der Metainduktivist: Ein spieltheoretischer Zugang zum Induktionsproblem. In R. Bluhm and C. Nimtz, editors, *Selected Papers Contributed to the Sections of GAP.5, Fünfter Internationaler Kongress der Gesellschaft für Analytische Philosophie, Bielefeld, 22–26 September 2003*, pages 243–257, Paderborn, 2004. Mentis.

G. Schurz. The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science*, 75(3):278–305, 2008.

G. Schurz. Meta-induction and social epistemology: computer simulations of prediction games. *Episteme*, 6(2):200–220, 2009.

G. Schurz. No free lunch theorem, inductive skepticism, and the optimality of meta-induction. *Philosophy of Science*, 84(4):825–839, 2017.

G. Schurz. Optimality justifications: new foundations for foundation-oriented epistemology. *Synthese*, 195(9):3877–3897, 2018.

G. Schurz. *Hume's Problem Solved: The Optimality of Meta-Induction*. 20xx. Manuscript in preparation.

G. Schurz and P. D. Thorn. The revenge of ecological rationality: Strategy-selection by meta-induction within changing environments. *Minds and Machines*, 26(1–2):31–59, 2016.

W. Sellars. Induction as vindication. *Philosophy of Science*, 31(3):197–231, 1964.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory To Algorithms*. Cambridge University Press, Cambridge, 2014.

B. Skyrms. On failing to vindicate induction. *Philosophy of Science*, 32(3):253–268, 1965.

B. Skyrms. *Choice and Chance: An Introduction to Inductive Logic*. Wadsworth, 4th edition, 2000.

T. F. Sterkenburg. *Universal Prediction: A Philosophical Investigation*. PhD Dissertation, University of Groningen, 2018.

T. F. Sterkenburg. The meta-inductive justification of induction: The pool of strategies. To appear in *Philosophy of Science*, 2019.

V. G. Vovk. Aggregating strategies. In M. Fulk and J. Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT90)*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

V. G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56:153–173, 1998.

V. G. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Munich Center for Mathematical Philosophy, LMU Munich
Email address: tom.sterkenburg@lmu.de