# No Free Lunch Theorem, Inductive Skepticism, and the Optimality of Meta-induction

Gerhard Schurz*†

The no free lunch theorem is a radicalized version of Hume's induction skepticism. It asserts that relative to a uniform probability distribution over all possible worlds, all computable prediction algorithms—whether 'clever' inductive or 'stupid' guessing methods (etc.)—have the same expected predictive success. This theorem seems to be in conflict with results about meta-induction. According to these results, certain meta-inductive prediction strategies may dominate other (non-meta-inductive) methods in their predictive success (in the long run). In this article this conflict is analyzed and dissolved, by means of probabilistic analysis and computer simulation.

**1. The Optimality of Meta-induction: A Solution to the Problem of Induction?** Schurz (2008) developed a new account to the problem of induction that is based on the optimality of meta-induction. The account agrees with Hume's skeptical insight that it is impossible to demonstrate a priori that induction is *reliable* in the sense that it is predictively more successful than random guessing. Such a demonstration is impossible without assuming that the actual world possesses a certain amount of regularity. Reichenbach (1949, sec. 91) argued that it is at least possible to demonstrate a priori that induction is *optimal*, that is, is the best what we can do for the purpose of predictive success. Results in formal learning show, however, that it is not possible to demonstrate optimality at the level of *object-induction*, that is, of induction applied to the task of predicting events in arbitrary possible worlds (cf. Skyrms 2000, sec. III.4). In contrast, what the account of *meta-induction* attempts to show

*To contact the author, please write to: Department of Philosophy (DCLPS), Heinrich Heine University Duesseldorf, Universitaetsstrasse 1, Geb. 24.52, 40225 Duesseldorf, Germany; e-mail: schurz@phil.uni-duesseldorf.de.

†This work was supported by the DFG (Deutsche Forschungsgemeinschaft), SPP 1516. For valuable help I am indebted to Paul Thorn and Ronald Ortner.

Philosophy of Science, 84 (December 2017) pp. 825–839. 0031-8248/2017/8405-0003$10.00
Copyright 2017 by the Philosophy of Science Association. All rights reserved.



https://doi.org/10.1086/693929 Published online by Cambridge University Press

is that induction is optimal if it is applied at the meta-level of competing prediction methods. The meta-inductive strategy tracks the success rate of all prediction methods whose predictions are *accessible* and predicts an optimal weighted average of the predictions of those methods that were most successful so far. Using results in the theory of regret-based online learning (Cesa-Bianchi and Lugosi 2006), Schurz (2008) proved that there exists a particular weighting method, called *attractivity-weighting*, which grants the meta-inductivist a predictive success rate that is in the long run at least as high as that of every other prediction method that is accessible to the meta-inductivist. Since the restriction to accessible methods is crucial for the optimality theorem, Schurz and Thorn (2016) call this kind of optimality *access-optimality*. Remarkably, the access-optimality of meta-induction holds in all possible worlds, even in 'chaotic' worlds in which the event frequencies and success rates do not converge against limits but are permanently changing in irregular ways, or in 'paranormal' worlds that host perfect clairvoyants or demonic deceivers.

Technically the account of meta-induction is based on the notion of a prediction game:

**Definition 1**. A *prediction game* is a pair $((e),\prod)$ consisting of

1. an infinite sequence $(e) := (e_1, e_2, \dots)$ of events $e_n$, coded by real numbers between 0 and 1, possibly rounded according to a finite accuracy. For example, (e) may be a sequence of daily weather conditions, football game results, or stock values. In what follows $Val \subseteq [0, 1]$ denotes the value space of possible events $e_n \in Val$. Each time n corresponds to one round of the game.
2. a finite set of prediction methods or 'players' $\prod = \{P_1, \dots, P_m, MI\}$ (in what follows we identify 'methods' with 'players'). In each round it is the task of each player to predict the next event of the event sequence. "MI" signifies the meta-inductivist and the other players are the 'non-MI-players' or 'candidate methods'. They may be real-life experts, virtual players implemented by computational algorithms, or even 'clairvoyants' who can see the future in 'paranormal' possible worlds. It is assumed that the predictions of the non-MI players are accessible to the meta-inductivist. Moreover, they are elements of a value space $Val_{pred}$ that may extend Val; that is, $Val \subseteq Val_{pred} \subseteq [0, 1]$.

The *predictive success rate* of a method P is defined by means of the following chain of definitions:

- $pred_n(P)$ is the prediction of player P for time n delivered at time $n - 1$,
- the deviation of the prediction $pred_n$ from the event $e_n$ is measured by a normalized loss function, $loss(pred_n, e_n) \in [0, 1]$,

https://doi.org/10.1086/693929 Published online by Cambridge University Press

- $\text{score}(\text{pred}_n, e_n) =_{\text{def}} 1 - \text{loss}(\text{pred}_n, e_n)$ is the score obtained by prediction $\text{pred}_n$ of event $e_n$,
- $\text{abs}_n(P) =_{\text{def}} \Sigma_{1 \le i \le n} \text{score}(\text{pred}_i(P), e_i)$ is the absolute success achieved by player P until time n, and
- $\text{suc}_n(P) =_{\text{def}} \text{abs}_n(P)/n$ is the success rate of player P at time n.

The natural loss function is defined as $|\text{pred}_n - e_n|$. The optimality theorem below holds for all *convex* loss functions, which means that the loss of a weighted average of two predictions is not greater than the weighted average of the losses of two predictions. In what follows we assume convex loss functions; they comprise a large variety of loss functions including all linear, polynomial, or exponential functions of the natural loss function.

'Possible worlds' are identified with prediction games. A special case are *binary* games whose events and predictions are elements of $\{0,1\}$. For binary games the natural loss function coincides with the zero-one loss: $\text{loss}_{1-0}(\text{pred}, e) = 0$ if $\text{pred} = e$, and otherwise $= 1$.

The simplest meta-inductive strategy is called "imitate-the-best" and predicts what the currently best non-MI player predicts. It is easy to see that this meta-inductive method cannot be universally access-optimal: its success rate breaks down when it plays against non-MI methods that are *deceivers*, which means that they lower their success rate as soon as their predictions are imitated by the meta-inductivist (cf. Schurz 2008, sec. 4). A realistic example is the prediction of stock values in a 'bubble economy': here the prediction that a given stock will yield a high rate of return leads many investors to put their money on this stock, and by doing so they cause it to crash. Nevertheless there exists a meta-inductive strategy that is provably universally optimal. This strategy is called *attractivity-weighted meta-induction* (wMI) and is defined as follows:

**Definition 2.** The predictions of wMI (attractivity-weighted meta-induction) are defined as

$$\text{pred}_{n+1}(\text{wMI}) =_{\text{def}} \frac{\sum_{1 \le i \le m} \text{at}_n(P_i) \times \text{pred}_{n+1}(P_i)}{\sum_{1 \le i \le m} \text{at}_n(P_i)},$$

where

- $\text{at}_n(P_i)$ is the attractivity of a player $P_i$ for wMI at a given time n, defined as $\text{at}_n(P_i) =_{\text{def}} \text{suc}_n(P_i) - \text{suc}_n(\text{wMI})$, if this expression is positive; else $\text{at}_n(P_i) = 0$, and
- if $n = 1$ or the denominator is 0, wMI's prediction is a random guess.

Attractivity $\text{at}_n(P_i)$ is also called wMI's *regret* compared to $P_i$ (at time n), and attractivity-weighted meta-induction is a case of regret-based learning.

https://doi.org/10.1086/693929 Published online by Cambridge University Press

Let "maxsuc$_n$" denote the non-MI-players' maximal success rate at time n. Then the optimality theorem for wMI asserts (see Schurz 2008, sec. 7, theorem 4):

> **Theorem 1.** (Universal access-optimality for wMI): For every prediction game $((e),\{P_1, \ldots, P_m, \text{wMI}\})$ the following theorems hold:
>
> 1.1. Short run: $(\forall\ n \geq 1:)\ \text{suc}_n(\text{wMI}) \geq \text{maxsuc}_n - \sqrt{m/n}$.
> 1.2. Long run: $\text{suc}_n(\text{wMI})$ (strictly) converges to the non-MI-players' maximal success for $n \to \infty$.

According to theorem 1.2 attractivity-weighted meta-induction is long-run optimal for all possible event sequences and sets of accessible prediction methods. The only proviso is that the set of accessible methods is finite, which is a realistic assumption for cognitively finite beings. In the short run, weighted meta-induction may suffer from a possible loss, compared to the leading player. This loss is caused by the fact that wMI must base her prediction of the next event on the past success rates of the candidate methods, and the hitherto most attractive methods may perform badly in the prediction of the next event. Fortunately theorem 1.1 states a worst case upper bound for this loss, $\sqrt{m/n}$, which is small if the number of competing methods, m, is small compared to the number of rounds, n, and which converges quickly to zero when n grows large.[1]

Theorem 1 applies to prediction games with real-valued as well as binary (or discrete) events. Even if the events are binary wMI's predictions are real valued (because proper weighted averages of zeroes and ones are real valued). How can the optimality result of theorem 1 be transferred to binary games whose predictions must be binary? There are two methods by which this can be done:

1. Randomization, rwMI (Cesa-Bianchi and Lugosi 2006, sec. 4.1): Here one assumes that rwMI predicts $e_n = 1$ with a probability (P) that equals the optimal real-valued prediction of wMI; that is, $P(\text{pred}_n(\text{rwMI}) = 1) = \text{pred}_n(\text{wMI})$. This method is not entirely general since it presupposes that the events are probabilistically independent from rwMI's choice of prediction.
2. Collective meta-induction, cwMI (Schurz 2008, sec. 8): Here a collective of meta-inductivists approximates real-valued predictions by the mean value of their binary predictions. Their mean predictive suc-

1. By using (more complicated) exponential attractivities this bound can be improved to (approximately) $1.77 \times \sqrt{\ln(m)/n}$ (Cesa-Bianchi and Lugosi 2006, 17, theorem 2.3).

https://doi.org/10.1086/693929 Published online by Cambridge University Press

cess rate approximates provably the success rate of the optimal method wMI. Assuming that the cwMIs are cooperators and share their success, every individual member of the collective is predictively optimal.

Theorem 1 establishes the following a priori justification of attractivity-weighted meta-induction: in all environments it is reasonable—in addition to searching for good object-level methods—to apply the strategy wMI, as this can only improve but not worsen one's success in the long run. Note that by itself this justification does not entail anything about the rationality of object-level induction: it may well be that we live in a world in which a method different from object-induction is predictively superior. However, the a priori justification of meta-induction gives us the following a posteriori justification of object-induction: to the extent that so far object-inductive prediction methods were more successful than all accessible noninductive (object-level) prediction methods, it is meta-inductively justified to continue favoring object-inductive prediction methods in the future.

Theorem 1 asserts the optimality but not the dominance (in the long run) of attractivity-based meta-induction. Thus, there may exists other methods, different from wMI, that are likewise long-run optimal. In fact one can prove that there are certain variants of wMI that are long-run optimal and have short-run advantages in certain environments and disadvantages in others. So wMI cannot be universally long-run dominant. Nevertheless, the following restricted dominance result for wMI follows from theorem 1:

**Theorem 2.** (Dominance for wMI):

2.1. wMI dominates every prediction method that is not universally long-run optimal. In other words, for every such method M there is a prediction game containing wMI and M in which wMI's long-run success rate exceeds that of M.
2.2. Not universally long-run optimal are, for example, all independent nonclairvoyant methods, that is, methods that can learn only from observations of past events but not from the predictions of other methods.

*Proof of theorem 2* Theorem 2.1 is an immediate consequence of theorem 1 and the definition of "optimality." The proof of theorem 2.2 goes as follows: let M be an independent method based on a function f that maps each n-tuple of past events $(e_1, \dots, e_n) \in \mathrm{Val}^n$ into a prediction $\mathrm{pred}_{n+1} \in \mathrm{Val}$. We define an M-adversarial event sequence $(e')$ as follows: $e'_1 = 0.5$, and $e'_{n+1} = 1$ if $f(e'_1, \dots, e'_n) \leq 0.5$; else, $e'_{n+1} = 0$. Moreover, we identify the predictions of the perfect $(e')$-forecaster $M'$ with the so-defined se-

https://doi.org/10.1086/693929 Published online by Cambridge University Press

quence; that is, $\text{pred}_n(M') = e'_n$ (note that if f is computable, M′ is too). In the prediction game $((e'),\{M,M',wMI\})$ the success rate of M can never exceed 1/2, that of M′ is always 1, and that of wMI converges to 1 (by theorem 1). So M is not universally long-run optimal, which proves theorem 2.2. QED

Theorem 2 is crucial for the next sections in which we confront the optimality of meta-induction with the no free lunch theorem.

**2. Radical Inductive Skepticism: The No Free Lunch Theorem.** The (in)famous no free lunch (NFL) theorem is a radicalized version of Hume's inductive skepticism for theoretical computer science.[2] The theorem applies to prediction methods that can be represented as computable functions from past observations to predictions, so-called learning algorithms (thus, clairvoyance is excluded). The theorem is often expressed by the assertion that for each pair of prediction methods, the number—or in the infinite case, the probability—of possible worlds (event sequences) in which the first method outperforms the second is precisely equal to the number (or probability) of worlds in which the second method outperforms the first. We call this assertion the *strong* NFL theorem, because it presupposes a 'homogeneous' loss function:

**Theorem 3.** (Strong NFL theorem; Wolpert 1996, 1354–55, theorems 1, 3): For every possible loss value c, the probability of worlds in which the prediction method leads to a loss of c is the same for all possible prediction methods, provided one assumes

*a*) a *state-uniform* prior probability distribution, that is, a uniform distribution over all possible event sequences (or states of the world), and
*b*) a homogeneous loss function, in the sense that for all possible loss values c, the number of possible events $e \in \text{Val}$ for which a prediction pred leads to a loss of c is the same for all possible predictions $\text{pred} \in \text{Val}_{\text{pred}}$.

The requirement of a homogeneous loss function is very strong: it is only satisfied if events and predictions are binary or, more generally, if they both discrete with $\text{Val} = \text{Val}_{\text{pred}}$ and a zero-one loss function. Under this assumption homogeneity is obvious: if the value space has k elements, then for every

2. The most general version of the NFL is found in Wolpert (1996); see also Schaffer (1994) and Rao, Gordon, and Spears (1995).

https://doi.org/10.1086/693929 Published online by Cambridge University Press

pred ∈ Val the number of possible events e ∈ Val that lead to a loss of 1 is obviously k − 1, and the number of events that lead to a loss of 0 is one. In contrast, in prediction games with real-valued predictions the homogeneity requirement fails. In the binary case, for example, the number of events that lead to a loss of 1 is one for the two predictions pred = 1 and pred = 0 but zero for the prediction pred = 0.5.

Homogeneous loss functions are a clear restriction of the strong NFL theorem. Real-valued predictions make sense not only for real-valued but also for binary events (e.g., by predicting their probabilities). Moreover (as explained) they can be implemented in binary games by randomization over binary predictions or by a cooperative collective of binary forecasters. There is, however, a weak version of the NFL theorem (mentioned by Wolpert 1996, 1354) that applies to real-valued predictions over binary or discrete events and assumes what we call a "weakly homogeneous" loss function:

> **Theorem 4.** (Weak NFL theorem; Wolpert 1996, 1354): The probabilistically expected success of every possible prediction method is equal to the expected success of random guessing or of every other prediction method, provided one assumes
>
> *a*) a state-uniform prior probability distribution, and
> *b*) a weakly homogeneous loss function, in the sense that for every possible prediction pred the sum of pred's losses over all possible events e ∈ Val is the same ($\forall$ pred ∈ $\mathrm{Val}_{\mathrm{pred}}$: $\Sigma_{e \in \mathrm{Val}} \mathrm{loss}(\mathrm{pred}, e) =$ a constant $c^*$).

For binary events with real-valued predictions and a natural loss function, weak homogeneity is satisfied since for every prediction pred ∈ [0, 1], loss(pred, 1) + loss(pred, 0) = 1 − pred + pred = 1.

For prediction games with real-valued events, most loss functions (including all convex ones) are not even weakly homogeneous. Here "free lunches" are possible in the sense that not all prediction methods have the same expected success, relative to a state-uniform probability distribution.

In this article we focus on prediction games with discrete events and real-valued predictions, to which the weak NFL theorem applies. Wolpert (1996) proves the NFL theorems in the framework of learning algorithms that map training sets into predictions of test items. A prediction game can be considered as an iteration of the procedure of selecting the first n events as a training set and the next event as a test item, that is, the test item of the nth round is added to the training set of the next round. Because a state-uniform prior induces a state-uniform conditional distribution $P(e_{n+1}|(e_1, \ldots, e_n))$, the NFL theorem applies straightforwardly to the (expected) scores in each round

https://doi.org/10.1086/693929 Published online by Cambridge University Press

of a prediction game and thus (by the linearity of expectation values) to the (expected) score sums in the entire game.[3]

Theorem 4 implies that every possible prediction method—be it an intelligent inductive one, a crazy anti-inductive one, or a stupid one that always predicts the same value—has the same expected predictive success relative to a state-uniform prior distribution. For all induction-friendly philosophical programs, including the program of meta-induction, this result seems to be devastating. How is it possible? In what follows we give a brief explanation of theorem 4 in terms that are philosophically more familiar than Wolpert's "extended Bayesian framework."

Wolpert's NFL theorem is a far-reaching generalization of a straightforward result about the prediction of binary sequences. For this application the NFL theorem rests on the following fact: however a prediction function f, with $\text{pred}_{n+1} = f((e_1, \ldots, e_n)) \in \{0, 1\}$, is defined, there are as many sequences of a given length $k > n$ extending $(e_1, \ldots, e_n)$ that verify f's prediction $\text{pred}_{n+1}$ as there are sequences that falsify it. Thus, by attaching an equal probability to every possible sequence, the expected score of each prediction function will be $1/2$. The same observation is the content of an (in)famous result in probability theory that can be found (among other works) in Carnap (1950, 564–66) or Howson and Urbach (1996, 64–66). The result can be expressed as follows:

> **Theorem 5.** (Carnap 1950, 564–66): Let P be a state-uniform prior probability (density) distribution over (the Borel algebra over) the set of all infinite binary sequences, $\{0,1\}^{\omega}$.[4] Then P has the following two 'radically noninductive' properties:
>
> *a*) P assigns the same conditional probability to each event $e_n \in \{0, 1\}$ independently of the preceding events $(e_1, \ldots, e_{n-1})$ of the sequence. Thus, P is an independent identical distribution with $P(1) = P(0) = 1/2$.
>
> *b*) P assigns a probability of 1 to the class of sequences with a limiting frequency of $1/2$ and a probability of 0 to all other possible limiting frequencies; this follows from *a* by the strong law of large numbers.

3. For the application of the NFL theorem to prediction games, the state uniformity of P is essential; generalizations of the NFL theorem to certain nonuniform distributions (cf. Rao et al. 1995) apply to 'one-shot' generalizations but not to prediction games.

4. Probability P yields Carnap's confirmation function $c^{\dagger}$. Technically, $\{0,1\}^{\omega}$ is represented by the interval [0,1] of real numbers in binary representation (see fig. 1). Probability P over the Borel algebra Bo([0,1]) is defined by the integrals of an assumed density function $D_P$ over [0,1].

https://doi.org/10.1086/693929 Published online by Cambridge University Press

**3. No Free Lunch and Meta-induction—a Conflict?** We now turn to the relation between the weak NFL theorem and theorem 2 about meta-induction. The NFL theorem applies not only to object-level prediction methods but also to all meta-strategies, given that they are applied to a fixed set of independent prediction methods—for the reason that every combination of a finite number of prediction algorithms is itself a prediction algorithm. So the puzzling question arises: If the NFL theorem is true, how can it be that attractivity-weighted meta-induction, when applied to a fixed set of independent prediction algorithms, is dominant in comparison to certain other methods, as stated in theorem 2? Is this a contradiction?

Our answer to this question in regard to the long-run perspective can be summarized as follows: no, the contradiction is only apparent. It is indeed true that there exist many wMI-accessible methods whose predictive success rate is (in the long run) strictly smaller than that of wMI in some worlds (event sequences) and never greater than that of wMI in any world—let us call these methods $M_{inf}$ (for "inferior").[5] Nevertheless, the state-uniform expectation values of the success rates of wMI and $M_{inf}$ are equal because a state-uniform distribution assigns a probability of 0 to all worlds in which wMI dominates $M_{inf}$, so these worlds do not affect the probabilistic expectation value. Put into a nutshell, for weakly homogeneous loss functions free lunches exist but receive a probability of 0 under a state-uniform prior.

Let us elaborate on this connection. The major difference between the account of meta-induction and Wolpert's extended Bayesian account is this: while the former account is independent from any assumed prior distribution over possible event sequences, Wolpert's result depends on a particular prior distribution, the state-uniform distribution. Wolpert seems to assume that this distribution is epistemically privileged. Reasonable doubts can be raised here because the state-uniform distribution is induction hostile. A proponent of this distribution is a priori certain that the binary event sequence she is going to predict (*a*) has a limiting frequency of $1/2$ and (*b*) is noncomputable. Fact *a* follows from theorem 5, and *b* from the fact that there are uncountably many sequences but only countably many computable ones. However, the event sequences for which an intelligent prediction method can be better than random guessing or any other 'stupid' method are precisely those event sequences that do not fall into the intersection of classes *a* or *b*. Let us make this point explicit: for random sequences with a limiting frequency of $1/2$, all combinations of independent methods must have the same success rate as random guessing (i.e., $1/2$). The only possibility for these sequences

5. Generally speaking possible worlds are identified with prediction games. But in the given context, we assume a fixed set of prediction methods, whence possible worlds can be identified with event sequences.

https://doi.org/10.1086/693929 Published online by Cambridge University Press

to be predictable is that they are computable by an internal regularity, but this possibility has probability 0, too.

In conclusion, proponents of a state-uniform prior distribution are strongly biased: they are a priori certain that the world is irregular so that induction cannot have any chance. We suppose that adherents of a more induction-friendly view, for example, Bayesians in the ordinary (not Wolpertian) sense, will regard a state-uniform prior distribution as highly "unnatural." Instead of a state-uniform distribution, they typically prefer a uniform distribution over all possible limiting frequencies; we call such a distribution a *frequency-uniform* distribution. It is well known that frequency-uniform distributions are highly induction friendly: from them one can derive Laplace's rule of induction,

$$P\left(e_{n+1} = 1 \,\middle|\, f_n(1) = \frac{k}{n}\right) = \frac{k+1}{n+2},$$

where "$f_n(1)$" denotes the frequency of ones among the first n events (cf. Carnap 1950, 568). In computer science, Laplace's rule has been generalized by Solomonoff (1964, sec. 4.1), who proved that if the prior probability of a sequence is inversely proportional to its *algorithmic complexity*, then Laplace's rule of induction is valid.

The precise relation between prior distributions over the space of possible infinite sequences and corresponding distributions over the space of possible limiting frequencies (or classes of sequences with the same frequency) is displayed in figures 1 and 2. As usual, infinite 0–1 sequences are represented as real numbers between 0 and 1 in binary representation (e.g., 0.0110. . .) and ordered according to their numerical size. In this way, the state-uniform distribution over possible sequences is represented as a uniform density over the interval [0,1]. Figure 1 presents the transformation of this distribution into the corresponding distribution over possible limiting

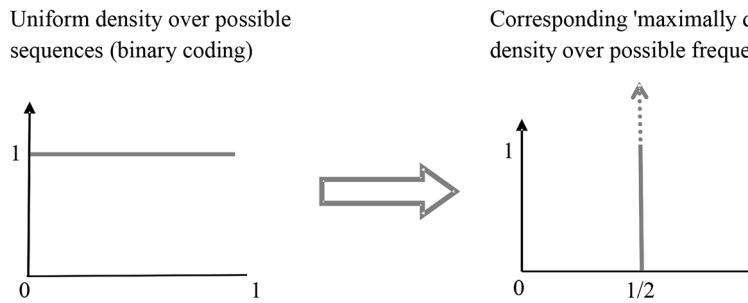



Figure 1. Transformation of a state-uniform distribution into the corresponding distribution over frequencies.

https://doi.org/10.1086/693929 Published online by Cambridge University Press

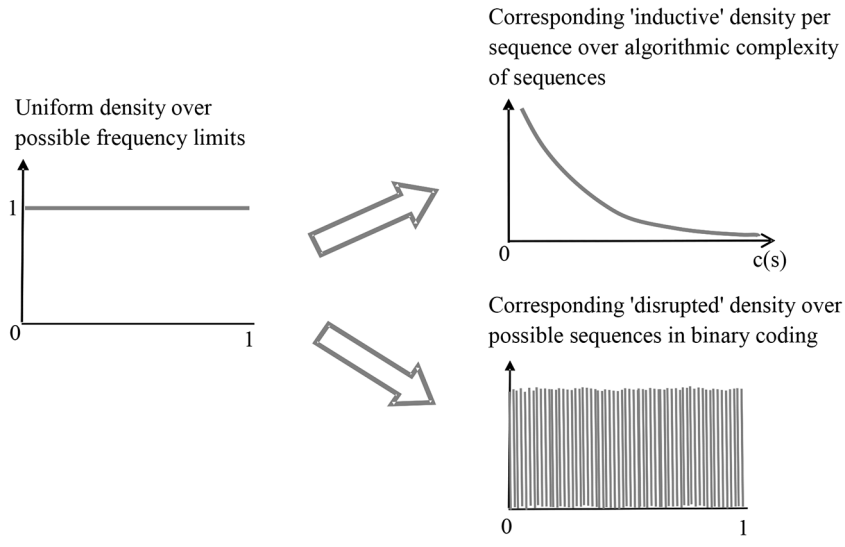



Figure 2. Transformation of a frequency-uniform distribution into the corresponding distribution over the algorithmic complexity of sequences (*top*) and over sequences in binary coding (*bottom*).

frequencies, with the result that a uniform distribution over [0,1] viewed as space of sequences is transformed into a maximally dogmatic distribution (an infinite density peak) over [0,1] viewed as space of frequency limits.

Figure 2 illustrates the inverse transformation. The upper part of figure 2 displays Solomonoff's (1964) result, which states that the frequency-uniform probability of a (finite or infinite) sequence decreases exponentially with its algorithmic complexity c(s): $P(s) \sim 2^{-c(s)}$. Thus, sequences with lower complexity have a higher frequency-uniform probability than those with high complexity. In conclusion, a frequency-uniform distribution is strongly biased in regard to less complex (more regular) sequences. The lower part of figure 2 shows what happens to a frequency-uniform distribution over [0,1] when it is transformed into a distribution over [0,1] viewed as space of possible sequences. The resulting distribution becomes noncontinuous and disrupted: in every finite interval $I \subseteq [0, 1]$ it increases infinitely often to a maximal value and falls back to a minimal value.[6] It follows that a state-uniform

6. To see this, let $r_1$ and $r_2$ ($r_2 > r_1$) be two infinite sequences represented as binary real numbers $r_1$ = "0.00 . . . (n times zero)11 . . . (one forever)" and $r_2$ = "0.00 . . . (n − 1 times zero)11 . . . (one forever)." Their complexity is minimal. The class of sequences lying between $r_1$ and $r_2$ contains sequences of all complexities between a minimal complexity (which increases with n) and the maximal complexity (which is possessed by sequences with frequency limit 1/2 and is infinite for infinite sequences). So, by Solo-

https://doi.org/10.1086/693929 Published online by Cambridge University Press

prior distribution makes Bayesian convergence results impossible because all these results presuppose a (not necessarily uniform but) continuous prior distribution over the possible frequencies (Earman 1992, 141ff.). Thus, "outwashing of priors" is impossible for state-uniform prior distributions.

So which prior distributions are more natural, state-uniform ones or frequency-uniform ones? In our eyes, this question has no reasonable answer because all prior distributions are subjective and biased in some respect. We regard it as a great advantage of the optimality of meta-induction that it holds regardless of any assumed prior probability distribution. For a frequency-uniform prior distribution the probability of worlds in which meta-induction dominates random guessing is close to 1. For a state-uniform prior the probability of worlds in which meta-induction dominates random guessing is 0. Nevertheless, infinitely many such worlds exist, and it is precisely in these worlds that intelligent prediction methods can have a chance at all. We should certainly not exclude these induction-friendly worlds from the start by assigning a probability of 0 to them. This concludes our discussion of the relation between meta-induction and the NFL theorem within the perspective of the long run.

**4. No Free Lunch and Meta-induction in the Short-Run Perspective.** The discussion of Wolpert's theorem within the perspective of the short run is more intricate. Recall that for finite sequences the advantage of meta-induction comes at a certain cost, which vanishes in the long run but is non-negligible for short sequences. Table 1 presents the result of a computer simulation of all possible prediction games with a length of 20 rounds, with binary events, three independent prediction methods, and wMI.[7] The considered independent methods were

- majority induction, M-I, which always predicts the event that so far has been in the majority and is 0.5 in the case of ties (i.e., $\text{pred}_{n+1} = 1/0.5/0$ if and only if $f_n(1) > / = / < 0.5$, respectively),
- majority anti-induction, M-AI, which predicts the opposite of M-I (i.e., $\text{pred}_{n+1} = 0/0.5/1$ if and only if $f_n(1) > / = / < 0.5$, respectively),
- averaging, Av, which always predicts 0.5.

Table 1 displays the frequencies of sequences for which the absolute success of a prediction method lies in a certain interval, with [0,1) being the low-

monoff's result, the density of the sequences in the interval $[r_1,r_2]$ drops from a maximal density to a minimal density (which is 0 for infinite sequences) and climbs back to a maximal density. Since this holds for every arbitrary small interval, the claim follows.
7. Computer simulations were performed by Paul Thorn.

https://doi.org/10.1086/693929 Published online by Cambridge University Press

TABLE 1. Computer Simulation of M-I, M-AI, Av, and wMI in All ($2^{20}$) Binary Sequences with 20 Rounds

| Absolute Success Interval | M-I | M-AI | Av | wMI |
|---|---|---|---|---|
| [0,1) | 0 | .000 | 0 | 0 |
| [1,2) | 0 | .003 | 0 | 0 |
| [2,3) | 0 | .029 | 0 | 0 |
| [3,4) | 0 | .159 | 0 | 0 |
| [4,5) | 0 | .618 | 0 | 0 |
| [5,6) | .537 | 1.824 | 0 | 0 |
| [6,7) | 3.540 | 4.254 | 0 | 0 |
| [7,8) | 9.579 | 8.035 | 0 | 0 |
| [8,9) | 15.622 | 12.476 | 0 | 36.491 |
| [9,10) | 18.346 | 16.065 | 0 | 23.472 |
| [10,11) | 17.915 | 18.157 | 100.000 | 14.835 |
| [11,12) | 15.046 | 17.510 | 0 | 11.880 |
| [12,13) | 10.266 | 12.854 | 0 | 7.469 |
| [13,14) | 5.635 | 6.305 | 0 | 3.595 |
| [14,15) | 2.448 | 1.611 | 0 | 1.513 |
| [15,16) | .821 | .098 | 0 | .560 |
| [16,17) | .204 | 0 | 0 | .153 |
| [17,18) | .035 | 0 | 0 | .029 |
| [18,19) | .004 | 0 | 0 | .003 |
| [19,20] | 0 | 0 | 0 | 0 |
| State uniform average | 10 | 10 | 10 | 10 |

Note.—In lines beginning with an interval, numbers represent percentage of sequences for which absolute successes in the intervals in the first column have been reached.

est and [19,20] the highest possible success interval. In accordance with the weak NFL theorem, one sees that the average success is the same for all four methods. Nevertheless, the frequency distributions over classes of sequences in which these methods reach certain success levels is remarkably different. The averaging method predicts 0.5 always and earns a sum of scores of 10 in all possible sequences. The object-inductive method M-I reaches a high success level in more worlds than the anti-inductive method M-AI (symmetrically, M-AI attains a low success level in more worlds than M-I). In compensation, the number of worlds in which the anti-inductive method does just a little better than averaging is significantly higher than the corresponding number of worlds for the inductive method.

On the basis of these results, we obtain a justification of object-induction and of meta-induction even within the induction-hostile perspective of a state-uniform prior distribution for short-run sequences. One can reasonably argue that what counts is to reach high success in those environments that allow for high success. This is what independent inductive methods do. At the same time one should protect oneself against low successes—this is what cautious methods of the type "averaging" do. The advantage of wMI meta-induction

https://doi.org/10.1086/693929 Published online by Cambridge University Press

is that it combines both—reaching high successes where it is possible (inspect the intervals [12,13)–[19,20]) and at the same time avoiding low successes (inspect the intervals [8,9) and [9,10)). Thus, wMI achieves "the best of both worlds." This, however, goes on the cost of a certain short-run loss (inspect the intervals [10,11) and [11,12)).

**5. Conclusion.** In this article we confronted the optimality of meta-induction with the NFL theorem. We demonstrated that the apparent conflict between these two results disappears when one considers that the NFL theorem assumes a state-uniform prior distribution over the set of all (binary) event sequences. This distribution assigns a probability of 0 to all infinite sequences that exhibit some sort of regularity that an intelligent prediction method could exploit. Short sequences were investigated by means of a computer simulation of all possible sequences of length 20. The result shows that in spite of having an equal expected predictive success, different prediction methods differ significantly in the frequency with which they reach certain success levels. Meta-induction turns out to offer the best combination of two abilities: exploiting regular sequences and avoiding losses in irregular sequences.

We emphasize that this characterization of the advantage of meta-induction holds for the induction-hostile state-uniform prior distribution. If one switches to a frequency-uniform prior distribution, the explained advantages of object- and meta-induction get amplified. Now M-I and wMI have highest predictive success in all classes of sequences whose frequencies are in the intervals [0,0.1), . . . , [0.3,0.4) and [0.6,0.7), . . . , [0.9,1]: wMI suffers from a small loss compared to M-I in these frequency intervals. In the frequency intervals [0.4,0.5) and [0.5,0.6) the picture is reversed: here M-AI and Av are more successful than M-I; wMI suffers from a small loss compared to M-AI and Av but is more successful than M-I. Because of space limitations we abstain from presenting the details.

## REFERENCES

Carnap, Rudolf. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.
Cesa-Bianchi, Nicolo, and Gabor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge: Cambridge University Press.
Earman, John. 1992. *Bayes or Bust?* Cambridge, MA: MIT Press.
Howson, Colin, and Peter Urbach. 1996. *Scientific Reasoning: The Bayesian Approach*. 2nd ed. Chicago: Open Court.
Rao, R. Bharat, Diana Gordon, and William Spears. 1995. "For Every Generalization Action, Is There Really an Equal and Opposite Reaction?" In *Machine Learning: Proceedings of ICML 1995*, ed. A. Preditis and S. J. Russell, 471–79. Burlington, MA: Morgan Kaufmann.
Reichenbach, Hans. 1949. *The Theory of Probability*. Berkeley: University of California Press.
Schaffer, Cullen. 1994. "A Conservation Law for Generalization Performance." In *Machine Learning: Proceedings of ICML 1994*, ed. W. W. Cohen and H. Hirsh, 259–65. Burlington, MA: Morgan Kaufmann.
Schurz, Gerhard. 2008. "The Meta-inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem." *Philosophy of Science* 75:278–305.

https://doi.org/10.1086/693929 Published online by Cambridge University Press

Schurz, Gerhard, and Paul Thorn. 2016. "The Revenge of Ecological Rationality: Strategy-Selection by Meta-induction." *Minds and Machines* 26 (1): 31–59.
Skyrms, Brian. 2000. *Choice and Chance*. 4th ed. Belmont, CA: Wadsworth.
Solomonoff, Ray J. 1964. "A Formal Theory of Inductive Inference." Pts. 1 and 2. *Information and Control* 7:1–22, 224–54.
Wolpert, David H. 1996. "The Lack of A Priori Distinctions between Learning Algorithms." *Neural Computation* 8 (7): 1341–90.

https://doi.org/10.1086/693929 Published online by Cambridge University Press