Meta-induction as a solution to the no free lunch theorem: reply to Wolpert

Gerhard Schurz (Heinrich Heine University Duesseldorf, Germany)

Appears as sec. 4 of my paper: "In Search for Optimal Epistemic Methods: New Insights About Meta-induction", ***Journal for General Philosophy of Science*** 2023.
doi.org/10.1007/s10838-023-09649-2

Wolpert claims in the abstract and in sec. 4 of his paper that my account would favor the induction-friendly frequency-uniform prior distribution. Let me start this section by emphasizing that this claim is wrong. On the contrary, in several passages in Schurz (2019) it is emphasized that meta-induction is not bound to any particular prior distribution (e.g. on pp. 71f., 167, 240-244). Rather, what I object to Wolpert's no free lunch (NFL) theorem is that this theorem rests on a particular prior, namely the induction-hostile state-uniform prior. Although the justification of meta-induction works even for the state-uniform prior, this justification becomes much stronger if one allows for different possible priors that are evaluated and aggregated by probabilistic meta-induction, including induction-friendly as well as induction-hostile priors. But nowhere in my book do I express a preference for frequency-uniform priors, and I wonder how Wolpert came to this misunderstanding.

Wolpert defends his account against my objection that the NFL theorem for predictions depends on a state-uniform prior, by presenting versions of this theorem that apparently do not assume a state-uniform prior. The goal of this section is to demonstrate that in fact these versions *do* assume a state-uniform prior, at least implicitly, by the consideration of (unweighted) sums or averages over all possibilities.

Wolpert's paper starts with a nice introduction presenting a game-strategy devised by Parrondo as an early example of a strategy of meta-induction, or online learning under expert advice (OLEA), as it is called in machine learning. In Parrondo's setting, methods are represented by sequences of bits of their payoffs, and a simplified version of Parrondo's strategy, call it P, imitates the prediction (or action) of the method that has highest cumulated payoff. Obviously, P is a version of ITB. Wolpert explains why

P is a good strategy, but it should be added that ITB is not universally optimal.

In sec. 2 Wolpert turns to the NFL theorems. They apply only to prediction methods that are *non-clairvoyant*, in the sense that the total information about the past events and success rates screens off the next event from its prediction – which is equation (3) in sec. 2 of Wolpert's paper. In sec. 2 (below equation (5)) Wolpert presents two versions of NFL theorems that are only inessentially different. Both versions compare the sum or average of the loss or cost of prediction methods *over all possible event sequences* (or states of the world) f, with the result that this cost sum or average cost is the same for all methods. There is a second and more important distinction, that between a strong and a weak variant of the NFL theorem. The *strong variant* of the NFL theorems is presented by Wolpert. This variant presupposes a *homogeneous* loss function in the sense of Wolpert (1996, 1349) – which is arguably a too strong condition on loss functions – while the weak NFL theorem assumes a merely weakly homogeneous loss function (see below).

Let C be the set of all possible losses resp. "one-shot" costs c, i.e. the possible differences between a prediction and an event (formally $C = \{c: \exists pred \in Val_{pred} \exists e \in Val: c = loss(pred,e)\}$). The strong variant of the NFL theorem (in both of Wolpert's versions) applies to each possible cost value $c \in C$ and asserts, in simplified worlds, that the probability of having loss c averaged over all environments is the same for all non-clairvoyant methods. More precisely, version 1 of Wolpert's NFL theorem asserts that for all $c \in C$, the sum of the probabilities of a method's attaining cost c in world state f, summed over all possible f's (conditional on data of size m) is the same for all methods (note that Wolpert's variable C_{OTS} ranges over these possible c's).[1] Wolpert's version 2 asserts that for all $c \in C$, the probability of a method attaining cost c in world state f (conditional on a data sequence d) is the same for all methods, given a state-uniform probability distribution P(f) over the f's. Now, Wolpert says that a

[1] We ignore here Wolpert's probability $\pi(q)$ of choosing the predicted event q, because q is fixed.

> "secondary implication of the NFL theorems is that if it so happens that you assume/believe that P(f) is uniform, then the average over f's used in the NFL for search theorem [= version 1, G.S.] is the same as P(f) in version 2".

I don't think this implication is "secondary" because *summing up* the probabilities of attaining cost c in f over all f's is essentially the same as averaging over these probabilities (since dividing their sum by their number gives the average) which is in turn essentially the same as calculating the overall probability of attaining cost c by a uniform prior distribution over the f's (since the average of these probabilities over all f's equals their expected probability according to a state-uniform prior over the f's).

The condition of *homogeneity* requires that for *every possible loss value* $c \in C$, the number of possible event values $e \in Val$ for which a given prediction pred leads to a loss of c is the same for all possible predictions $pred \in Val_{pred}$. Homogeneity is satisfied only for prediction games with a zero-one loss function, which gives a maximal loss of one if the prediction differs from the event and a zero-loss if the prediction equals the event (cf. Schurz 2019, def. 9-1, 326). Obviously homogeneous loss functions are unreasonable whenever predictions and/or events are graded. For example, the prediction "0.9" of the event "1" is better than the prediction "0.1" (since the distance between 0.9 and 1 is much smaller than that between 0.1 and 1), although for homogeneous loss functions both predictions are equally bad and attain a score of zero. Therefore Schurz (2019, sec. 9.1) and Schurz and Thorn (2022) concentrate their investigation on weakly homogeneous loss functions, that are mentioned by Wolpert (1996) in a small paragraph on p. 1354 ("More generally, for an even broader set of loss functions …"). A loss function is *weakly homogeneous* iff for each possible prediction pred, the *sum* (or average) of the losses over all possible events is the same. For binary games with real-valued predictions and absolute loss function, weak homogeneity is satisfied, since for every possible prediction $pred \in [0,1]$, $loss(pred,1) + loss(pred,0) = 1 - pred + pred = 1$ (Schurz 2019, def. 9.2, 327).

The *weak variant* of the NFL theorem makes the corresponding assertion not for

each cost value $c \in C$ separately, but merely for the sum or average of all cost values. In version 1 the weak NFL theorem says that the average cost over all possible event sequences f (conditional on data size m), defined as $\Sigma_{f,c} P(c|f,m) \cdot c$, is the same for all methods, and in version 2 it says that the probabilistically expected cost of a method (conditional on a data sequence d), defined as $\Sigma_f P(f) \cdot \Sigma_c P(c|d,f) \cdot c$, is the same for all methods according to a state-uniform distribution P(f) over the f's. Finally, note that loss-functions for real-valued events do not even satisfy the condition of weak homogeneity and Wolpert's version of the NFL theorem does not hold for real-valued events; however, a weaker version of the NFL theorem applies to them (as proved in ibid., prop. 9.3).

We now turn to Wolpert's arguments against my diagnosis that the NFL theorem for predictions depends on a state-uniform prior. These arguments and my objections to them apply equally to the strong and the weak variant of Wolpert's NFL theorems. In his first argument Wolpert (4th § after equ. (5))) says that

> "it must be emphasized that simply allowing [the prior – G.S] P(f) to be non-uniform, by itself, does not invalidate the NFL theorems",

and some lines later he says that the

> "NFL theorems do not assume that the universe is governed by a uniform prior in some objective sense."

Here we meet an important confusion that is also found in other machine learning texts (for example also in the paper quoted in Shogenji's sec. 5, as mentioned in my sec. 4), namely the following: When epistemologists speak of a *prior* probability they mean always a *subjective-epistemic* probability, i.e. a rational degree of belief, but *not* an objective probability (be it a statistical propensity or an objective single case chance). A 'prior' probability is defined as a distribution that one adopts, or should reasonably adopt, *prior to experience*; this notion *only makes sense* for an epistemic notion of

probability, but not for an objective one, because objective probabilities are *independent* from whether the subject has experience or not. When machine learners speak of an "objective prior", they just mean the *true unconditional* probability function over the possible states of a type of system; but this is entirely different from a prior in the epistemic sense. For this reason, Wolpert's accusation in sec. 4 (5th §) that "Schurz argues that one “should” adopt a single, specific prior … a uniform prior over frequencies" is *not only* incorrect because I never make any such assertion; in addition Wolpert's critique of this position – which is the position of Laplacean inductivists – is inappropriate because Wolpert assumes wrongly that the frequency-uniform prior is meant in the objective sense. Wolpert attempts to refute this misunderstood position by pointing out that "all of statistical physics is based on a uniform distribution over patterns, not over frequencies". Wolpert's misleading critique culminates in his devious diagnosis in the last paragraph of his paper that

> "Schurz’s proposal for a uniform prior over frequencies runs afoul of thousands (tens of thousands?) of previous experiments concerning the real, physical world".

This wrongs me twice: first because it is *not me* who assumes frequency-uniform distributions but Laplacean inductivists, and second I know quite well that distributions of microcanonical ensembles in thermodynamics are not frequency-uniform, as Wolpert rightly observes, but his observation is *besides the point*, because the frequency-uniform distributions to which induction-friendly probability theorists refer are meant as epistemic and not as objective probabilities.

Having clarified this confusion, let us get to Wolpert's second major argument against my diagnosis that the NFL theorems are based on a state-uniform epistemic prior. Namely, Wolpert writes (in the 2nd half of his sec. 2) that

> "allowing P(f)'s [i.e., the priors over event sequences – G.S.] to vary provides us

> with a new NFL theorem. In this new theorem, rather than compare the performance of two learning algorithms by uniformly averaging over all f's, we compare them by uniformly averaging over all P(f)'s".

As Wolpert continues, this uniform averaging results again in an NFL theorem (in both of his versions). This is no wonder – because a uniform average over all objective priors over the space of possible event sequences is just *a second order version of a uniform epistemic prior* that results in a uniform expected first order prior. For example, suppose that events are binary (0 or 1) and $p =_{def} p(1)$. Assuming a uniform (2nd order) prior density D(p) over all possible (1st order) priors $p \in [0,1]$, the resulting expected 1st order probability of the event 1 is given as $\int_0^1 p \cdot D(p)dp = |_0^1 p^2/2 = 1/2$, which is uniform at the 1st order level.

Thus, Wolpert's proposed method of averaging over possible prior distributions is just another version of a state-uniform prior distribution. In conclusion, Wolpert's attempts to escape the diagnosis that the NFL theorems for prediction depend on a state-uniform prior do not work, and his claim in the 3rd-last § of section 2 that this diagnosis is "simply wrong" seems to apply to itself.

Let us now briefly explain the solution to the challenge provided by the NFL theorems proposed by meta-induction. It follows from the dominance results for aMI (recall result (3) in sec. 1) that aMI enjoys free lunches over all methods that it dominates. How can that be in view of the NFL theorems – is this not a contradiction? My answer distinguishes between the long run and the short run perspective. In both perspectives, the answer is no. In regard to the long run perspective, the contradiction is only apparent, because the state-uniform probability distribution that Wolpert assumes assigns a probability of zero to all worlds (infinite event sequences) in which aMI dominates the inferior methods (cf. Schurz 2019, 70f., 241); so these worlds do not affect the probabilistic expectation value of the method's success. But although the state-uniform prior of worlds in which aMI meta-induction dominates inferior methods is zero, there are many – indeed uncountably many – such worlds and it is precisely in these worlds that

intelligent prediction methods can have any chance at all. We should not exclude these induction-friendly worlds from the start by assigning a probability of zero to them, which means that we should not restrict the epistemic priors to uniform priors.

Within the short-run perspective, the defense of meta-induction against the NFL challenge is more difficult, because here aMI suffers a small regret. Here we argue as follows. What counts are two things: (a) To reach *high* success in those environments which *allow* for high success by their intrinsic properties (uniformities). This is what independent inductive methods do. (b) To *protect* oneself against high losses (compared to average success) in induction-hostile environments. This is what cautious methods do, such as the method "averaging" that always predicts the average of all possible event values. The advantage of aMI is that it combines *both* accomplishments – reaching high success rates whenever possible and avoiding high losses; a demonstration of this fact by computer simulations is found in Schurz and Thorn (2022, sec. 5). In conclusion, aMI achieves 'the best of both worlds', although this comes at the cost of a small short-run regret of aMI that is acceptable given the mentioned advantages of aMI. In the case of discrete events with linear loss function, the NFL theorems imply that the state-uniform average of this short-run regret is the same for all methods; but the advantages (a) and (b) even hold under this induction-hostile assumption. For quadratic loss functions or more induction-friendly priors the short-run advantages of meta-induction get amplified (cf. Schurz and Thorn 2022, tables 3-8). Wolpert's notion of "head-to-head minimax distinctions" in his sec. 4 comes close to my proposed solution for the short run: the maximal regret of the methods is minimal for aMI, and yet aMI climbs to high successes in regular environments.

Finally a remark on Wolpert's nice construction of a competition between two meta-level algorithms in his sec. 3 – a meta-inductive method based on cross-validation, and a corresponding meta-anti-inductive method. Both meta-methods have access to the same candidate pool of methods; we abbreviate the two meta-level methods as MI and MAI. Schurz (2019, 93, 157) calls such competitions prediction *tournaments*, as opposed to prediction *games*, since in tournaments it is assumed that the preferred meta-

inductive method cannot access the competing meta-methods. Wolpert observes that for every prior P(f) over event sequences for which MI performs well, there exists corresponding prior P*(f) for which AMI performs equally well. This is certainly correct, but it does not affect the optimality result, because it assumes that MAI is not accessible to the method MI, while the optimality theorem applies only to accessible methods. As soon as MI is allowed to access AMI's predictions MI's success is granted to converge to AMI's success in environments in which AMI is optimal.