Optimal Probability Aggregation Based on Generalized Brier Scoring[*]

Christian J. Feldbacher-Escamilla   Gerhard Schurz

Summer 2020

Abstract

[717] In this paper we combine the theory of probability aggregation with results of machine learning theory concerning the optimality of predictions under expert advice. In probability aggregation theory several characterization results for linear aggregation exist. However, in linear aggregation weights are not fixed, but free parameters. We show how fixing such weights by success-based scores, a generalization of Brier scoring, allows for transferring the mentioned optimality results to the case of probability aggregation.

Keywords: probability aggregation, Brier score, meta-induction, machine learning, no regret algorithms

1 Introduction

Probability aggregation is the theory of how to adequately aggregate several probability distributions to a single one. It is an expansion of the theory of judgment aggregation that combined problems studied by social choice theory and logic as, e.g., problems of preference aggregation and questions of voting theory.

In the past, research in judgment aggregation centred around the disciplines of economics and political science (Arrow 1963), law (Kornhauser and Sager 1986), and philosophy (List and Pettit 2002). Recently, however, increasing work in judgment aggregation stems also from computer science and research on artificial intelligence (Grossi and Pigozzi 2014; Rossi, Venable, and Walsh 2011). In particular there is an increase of interest in judgment aggregation of researchers from artificial intelligence specifically in the domain of knowledge representation and multi-agent systems. One important connection between judgment aggregation and artificial intelligence is seen in the fact

[*][This text is published under the following bibliographical data: Feldbacher-Escamilla, Christian J. and Schurz, Gerhard (2020). "Optimal Probability Aggregation Based on Generalized Brier Scoring". In: *Annals of Mathematics and Artificial Intelligence* 88.7, pp. 717–734. DOI: 10.1007/s10472-019-09648-4. All page numbers of the published text are in square brackets. For more information about the underlying project, please have a look at http://cjf.escamilla.academia.name.]

that judgment aggregation has at its core the [718] task to resolve inconsistencies that might show up once one aggregates individually consistent opinion profiles to a single opinion profile. Since also a significant part of research in artificial intelligence and logic is about resolving inconsistencies, as, e.g., in nonmonotonic reasoning, belief revision, belief merging, paraconsistent logic, inconsistency debugging, etc., it seems quite natural to assume that new techniques studied in the theory of judgment aggregation might also be of some use for artificial intelligence research. An overview of possible applications and fruitful connections is provided, e.g., in (Dietrich et al. 2014).

Now, one fundamental problem of judgment aggregation consists in its theoretical underdetermination: Already in the 1980s several characterization results have been proven for families of probability aggregation rules (Genest and Zidek 1986). However, these characterizations leave some parameters still free and uninterpreted. In this paper we provide a new approach to fix these parameters. Following suggestions of the literature on scoring rules for probabilistic forecasts (Genest and McConway 1990, pp.57ff), we suggest to interpret the weights in a success-based way. By cashing out results on no regret algorithms for prediction under expert advice in another field of computer science, namely online machine learning, we show that fixing the parameters in a success-based way allows for optimal probability aggregation.

The structure of the paper is as follows: In the following section we briefly present the basics of the framework of probability aggregation we are interested in. Afterwards we indicate the main result of research on prediction under expert advice, i.e. meta-induction, employed by us. Then we implement this result into the aggregation framework and show how it allows for optimal probability aggregation. We conclude in the final section.

2 Linear Probability Aggregation

The theory of probability aggregation deals with the problem of how to aggregate a set of probability distributions $P_1, \ldots, P_n$. For the probability distributions we assume that there is a finite real numbered measurable value space $\{v_1, \ldots, v_k\}$, that each P_i ($1 \leq i \leq n$) assigns to each v_j ($1 \leq j \leq k$) a nonnegative value, and that these mappings sum up to one for all values, i.e.:

$$P_i(v_j) \geq 0 \text{ and } \sum_{j=1}^{k} P_i(v_j) = 1$$

Abstractly speaking, the question of probability aggregation is how to characterize a probability aggregation rule f which takes as input a set of n probability distributions $P_1, \ldots, P_n$ and generates as output a/*the* aggregated probability distribution P_{aggr}:

$$P_{aggr} = f(P_1, \ldots, P_n)$$

Usually, several constraints are put forward for such an aggregation rule. Quite common are the following three constraints:

(U) *Universal domain*: f allows as input any P that satisfies the laws of probability theory

(Z) *Zero Unanimity*: f preserves unanimous zero-assignments: For all values $v_j \in \{v_1, \ldots, v_k\}$: $P_{aggr}(v_j) = f(P_1, \ldots, P_n)(v_j) = 0$, if $P_1(v_j) = \cdots = P_n(v_j) = 0$

(I) *Irrelevance of Alternatives*: f aggregates value-wise: There is an f^* such that for all values $v_j \in \{v_1, \ldots, v_k\}$: $P_{aggr}(v_j) = f(P_1, \ldots, P_n)(v_j) = f^*(P_1(v_j), \ldots, P_n(v_j))$

These constraints on aggregation are typically justified as follows: (U) allows for considering any individual probability distribution which is consistent. No such probability [719] distribution should be excluded on *a priori* grounds. By this, a wide range of possible inputs should be covered. (Z) is a very weak constraint for lifting unanimous considerations from the individual level to the collective one: If all individuals within a setting agree on a probabilistic assessment of 0 for an event, then also the aggregated outcome should agree on this. Finally, (I) is often argued for by principles of informational parsimony and avoiding "strategic voting" and manipulation: If one is interested in the aggregated probability of a single value v_j, then, if the aggregation method satisfies the irrelevance of alternatives constraint, one can concentrate on grasping individual $P_i(v_j)$ only, and one needs not to consider or collect probabilistic information about all the other values. Furthermore, if $P_{aggr}(v_j)$ depends on the individual probabilities of some other value v_l: $P_i(v_l)$, then individuals might have some incentive to dishonestly report their probability estimations in order to favor one alternative over another—i.e. the input of aggregation is prone to manipulation. So, all in all, these constraints seem to be justified once one is interested in the *aggregated* probability distribution of a set of individual probability distributions.

Furthermore, since we aim at combining probability aggregation with a particular scoring rule, it is also worth mentioning that some of these constraints for probability aggregation have formal pendants in the domain of scoring rules. Universal domain (U) is also commonly assumed to hold for scoring rules since such rules ought to operate on the whole interval of probabilistic forecasts; and, more interestingly, irrelevance of alternatives (I) is the formal pendant to the so-called *locality*-property of scoring functions where the score for the value only depends on the probability of the value in question (e.g., for trivial reasons are all binary scores local).

As is discussed and shown in (Genest and Zidek 1986; Lehrer and Wagner 1981, chpt.6 theorem 6.4, resp. sect.3), these three conditions for probability aggregation presented above characterize the family of linear probability aggregation rules which have the form of a weighted arithmetic mean (given

$k \geq 3$):

$$P_{aggr} = \sum_{i=1}^{n} w_i \cdot P_i \tag{LIN}$$

$$\text{(where } w_i \geq 0 \text{ and } w_1 + \cdots + w_n = 1\text{)}$$

So, any aggregation method which satisfies constraints (U,Z,I) is a linear (LIN) aggregation method, and every linear aggregation method satisfies these constraints. It is clear that according to this characterization result, different interpretations of the weights allow for different specifications. In this sense, probability aggregation is still underdetermined by the constraints (U,Z,I). In this paper we aim at further determining (LIN) by putting forward constraints on the weights. In particular, we argue for interpreting the weights in a regret-based way, because such an interpretation allows for optimal probability aggregation. In the next section we present such an optimality result.

Before we prepare the ground for our application, it should be noted that the mentioned constraints which characterize linear weighted aggregation run against other, also important constraints. So, e.g., independence preservation (if all individuals consider v_i and v_j to be probabilistically independent, then they should be also considered to be independent according to the aggregated result—see, e.g., (Genest and Zidek 1986, condition 3.4 in sect.3)) or the so-called notion of being *externally Bayesian* (which allows for some kind of commutativity between aggregation and Bayesian learning—for details see, e.g., (Genest, McConway, and Schervish 1986)) cannot be reconciled with linear weighted aggregation. Rather, they are characteristic for another important family of pooling methods, namely geometric pooling. Since here we focus on linear aggregation based on generalized Brier scoring, we will not go into further detail with respect to alternative constraints, and we leave it at this. [720]

3 Optimality in an Expert Advice Setting

In online machine learning regret bounds of algorithms for making predictions under expert advice in repeated prediction settings are studied (Cesa-Bianchi and Lugosi 2006). The idea is to consider a series of events whose outcomes have to be predicted by so-called *experts* or *candidate* methods. Given these predictions the task is to construct a prediction algorithm that uses the candidate method's forecast as input and aims at approaching the predictive success of the best expert(s) in the setting, even if the best expert is changing in time in irregular ways. Since a prediction method under expert advice combines the expert's predictions by inductively projecting the observed regrets to the future, it is called a *meta-inductive method* (Schurz 2008). Regrets are relative to a candidate method, and are defined as the difference between the *cumulative loss* of the meta-inductive algorithm and that of the respective candidate method. If it is positive, then the meta-inductive method has higher cumulative loss than the candidate method, and hence the meta-inductive method *regrets* in hindsight to not have predicted in accordance with the candidate method. If it is negative, then the meta-inductivist's cumulative loss is lower than that of the candidate method, and hence the meta-inductive method has not to regret to have predicted not in accordance with the candidate method. Since usually the loss at a prediction round is considered to be bounded by $[0,1]$, the score of a prediction for a round is defined as 1 minus its loss. Hence, the cumulative score up to prediction round t is t minus the cumulative loss. And hence the regret of the meta-inductive prediction method with respect to a candidate method can be defined as the difference between the accumulated score of the candidate method and that of the meta-inductivist. Positive regret means that the candidate method has a higher cumulative score, and negative regret means that the candidate method has a lower cumulative score. The idea of a *no regret algorithm* or method is to have no (positive) regret in the long run, i.e. that regret grows only sublinearly.

Here are the details: The settings of online learning are so-called *prediction games* that have the following ingredients (Schurz 2008, notation adjusted):

- E is an infinite series of events consisting of variables $E_1, E_2, \ldots$ whose outcomes $val_1(E), val_2(E), \ldots$ are elements of the normalized interval $[0,1]$.
- $P_{1,t}, \ldots, P_{n,t}$ are the predictions of E_t (also elements of $[0,1]$) of all n candidate methods.
- $P_{mi,t}$ is the prediction of E_t of the meta-inductive algorithm under investigation.

As we have indicated above, the meta-inductive algorithm "cooks up" a prediction from the present predictions and past success rates of the candidate methods. In order to keep track of the success rate of a method i one identifies the score of i's prediction about event E_t with 1 minus the loss l of this prediction and then sums up all of its scores up to round t and divides by t (Schurz 2008, sect.1):

$$s_{i,t} = \frac{\sum\limits_{u=1}^{t} 1 - l(P_{i,u}, val_u(E))}{t}$$

The measure $s_{i,t}$ represents the average per-round success rate of candidate method i up to round t. The only assumptions we make about the loss function l are (i) that it is within $[0,1]$, and (ii) that it is *convex* in its first argument, i.e. that the loss of a weighted average of two predictions is lower or equal to the weighted average of the losses of these two predictions. Formally: $l(w \cdot x + (1-w) \cdot y, z) \leq w \cdot l(x,z) + (1-w) \cdot l(y,z)$ holds for all x, y, z and $w \in [0,1]$. We put forward convexity as a desideratum for three reasons: first, convexity is a general property satisfied by a wide range of loss functions that are studied [721] in the area of probability aggregation; this desideratum is satisfied in particular by one kind of loss function we are mainly interested here, namely the quadratic loss function; second, also online learning, the branch to which we want to link our approach of probability aggregation, focuses on convex loss functions (via convex optimization; cf. (Cesa-Bianchi and Lugosi 2006)). And third, there are also non-technical reasons for the adequacy of this desideratum; so, e.g., if one shares the intuition that aggregating or "mixing" or "blending" individual opinions to a collective one should be rewarded by some kind of minimal benefit, then, e.g., only convexity serves a more fundamental constraint of providing a guarantee for avoiding strict sub-optimality of the aggregated opinion.

Now, based on this measure for the success rate up to round t one can define a so-called *attractivity* measure which serves for defining the weights of the meta-inductive prediction method. The idea of such a measure is that the higher the past success of an attractive method, the higher is also its weight. Moreover, the attractivity measure cuts off those candidate methods that are not attractive, i.e., that have a lower average per-round success rate as the algorithm. In linear weighting this is necessary in order for a meta-inductive method to approach the best candidate methods in the setting. Figure 1 illustrates the idea behind *cutting off*.

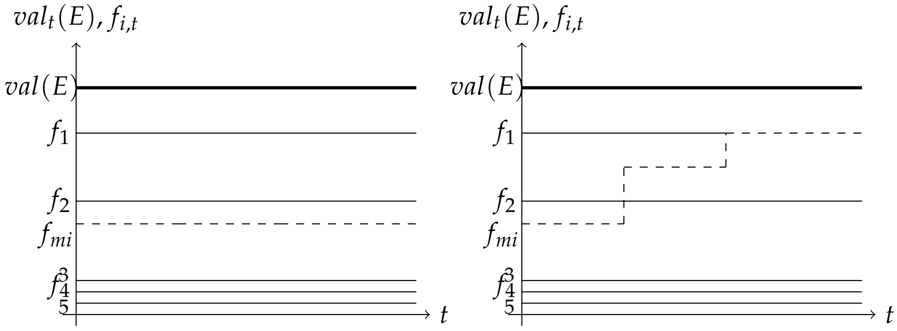

Figure 1: Example of taking success rates as weights without cutting off outperformed predictions (left) and with cutting off outperformed predictions (right): Ad left case: If the meta-inductive method (f_{mi}) just simply weights the predictions according to their success rates, the influence of predictions which are outperformed by it (here: f_3, f_4, f_5) might never vanish. This prevents that the meta-inductive method reaches the better, outperforming prediction methods (here: f_1, f_2). Ad right case: If the meta-inductive method cuts off the outperformed prediction methods, it reaches the better ones (first, by cutting off outperformed f_3, f_4, f_5, and then, by cutting off f_2 which is outperformed in the second round).

The weight of a candidate method P_i for the algorithm P_{mi} regarding event E_t is defined as follows (where $s_{mi,t}$ is the success per round of the meta-inductive method up to round t):

$$w_{i,t} = \frac{max(0, s_{i,t} - s_{mi,t})}{\sum\limits_{j=1}^{n} max(0, s_{j,t} - s_{mi,t})}$$

Candidate methods that are performing worse than P_{mi} get weight 0. If P_{mi} outperforms all candidate methods, then $s_{mi,t} \geq s_{i,t}$ for all $i \in \{1, \ldots, n\}$, and we stipulate $w_{i,t} = 1/n$. So, the weights are always positive and sum up to 1;

Based on these weights, we can define a weighted-average meta-inductive algorithm (MI) which weights the predictions of the candidate methods according to their attractivities. [722] Such an algorithm generates predictions by the method of linear (arithmetic) aggregation as follows (Cesa-Bianchi and Lugosi 2006; Schurz 2008, sect.2.1, resp. sect.7):

$$P_{mi,t+1} = \sum_{i=1}^{n} w_{i,t} \cdot P_{i,t+1} \tag{MI}$$

In case there are no attractive methods, but also at the very beginning (E_1) the algorithm's prediction consists of the mean of all predictions. Note that what we called here *attractivities* corresponds to positive *per-round regrets*.

The algorithm (MI) proves to be very powerful regarding the task of approaching the best candidate methods' per-round success rates: There are quite narrow bounds of P_{mi} regarding the worst-case per-round regret, i.e., the difference of its success rate compared to the success rate of the actually best candidate method. The basic result of the machine learning literature we want to employ in this paper is the following theorem on the upper bounds of the regret (Cesa-Bianchi and Lugosi 2006; Schurz 2008, sect.2.1f, resp. sect.7):

Theorem 1. *Given the underlying loss function l is convex it holds:*

$$s_{i,t} - s_{mi,t} \leq \sqrt{n/t} \quad \forall i \in \{1, \ldots, n\}$$

This theorem shows that (MI) is a no regret algorithm in the sense that:

$$\lim_{t \to \infty} s_{i,t} - s_{mi,t} \quad \leq \quad 0 \quad \forall i \in \{1, \ldots, n\}$$

So, the meta-inductive algorithm's success rate and that of the best performing candidate methods converge in the limit or the meta-inductivist even outperforms it. In this sense the meta-inductive algorithm is optimal. In the machine learning literature settings that allow for such a result are also known as *online learnable* (Shalev-Shwartz and Ben-David 2014).

The guaranteed performance of (MI) can be enhanced further by exponentially weighting the absolute regrets such that the upper bound of the per-round regret is $\sqrt{c \cdot \log(n)/t}$ with $c \geq .5$. Up to now the algorithm with the best known general upper bound is such an algorithm using exponentially absolute regret-weighting which guarantees such an upper bound with $c = 3.125$.

To design an algorithm which has the minimal upper bound that is achievable in principle, namely $\sqrt{\log(n)/2t}$ (Cesa-Bianchi and Lugosi 2006, p. 62, thrm. 3.7), is still an open task of online machine learning theory.

For our proposal the exact short-run bounds do not matter. What is relevant is that they allow for no regret in the long run. I.e., any way of cooking up success-based weights which allow for sublinear growth of regret (i.e. guaranteed decreasing per-round regret) serve the purpose of theoretically justifying this choice. In the next section we are going to utilize this result of meta-inductive optimality in order to fix the weights of linear probability aggregation and provide a rationale for doing so.

4 Optimal Probability Aggregation

Note that up to now the predictions in the prediction game were only about providing an estimate of *one* value within $[0,1]$. In this sense the predictions were not probabilistic, but deterministic. In the probabilistic case, each individual has to provide such an estimate for all possible values of $\{v_1, \ldots, v_n\}$ such that the estimations turn out to satisfy the probability constraint (of summing up to one). We can implement this by designing so-called *probabilistic prediction games*. [723]

In *probabilistic* prediction games each forecaster or candidate method identifies the predicted real value with its credence of the predicted event conditional on her information about the past. In the following part of this paper we are implementing the optimality result of the foregoing section into the framework of probability aggregation. In order to cash out the no regret optimality result of meta-induction presented above for probability aggregation we have to modify our framework a bit: It contains:

- Again, a series of events represented by random variables $E_1, E_2, \ldots$, but now the events do not have outcomes within $[0,1]$, but within a space of discrete (non-numerical), mutually disjoint and exhaustive values v_i, $\{v_1, \ldots, v_k\}$. In order to indicate which value a random variable took on at a specific round, we assume a valuation function *val* to be given by:

$$val_t(v_m) = \begin{cases} 1, & \text{if the value of } E_t \text{ is } v_m \\ 0, & \text{otherwise} \end{cases}$$

- Predictions are the credences of n candidate methods for each event variable E_t in the series, represented by probability distributions $P_1, \ldots, P_n$:

$$\forall\, t, i \in \{1, \ldots, n\} \sum_{m=1}^{k} P_{i,t}(v_m) = 1 \text{ and } P_{i,t}(v_m) \geq 0$$

  So, for each event, at each round, the candidate methods provide a full probability distribution about the outcome of the event in question.

- The meta-inductive algorithm P_{mi} is also represented by a probability distribution and defined as an arithmetically weighted average of the $P_1, \ldots, P_n$; details are presented below.

The attempt to expand the framework of prediction games introduced in the foregoing section to the probabilistic setting faces the problem that the predictions are real numbers, i.e. probabilities, but the event's values are not numbers but non-numeric mutually exclusive and exhaustive values $v_1, \ldots, v_k$. However, as we will show now, there is a possibility of applying the meta-inductive framework of prediction games to this case.

Since each of these values has two possible truth values, 0 and 1, we can score probabilistic predictions by comparing them with these truth values for each of the possible values. This means in effect that we mimic a prediction game about a random variable with k values $v_1, \ldots, v_k$ by launching k prediction games about k binary events, v_m versus not-v_m, in parallel.

We can define a measure for the predictive success regarding a value v_m as follows:

$$s_{i,t}(v_m) = \frac{\sum_{u=1}^{t} 1 - l(P_{i,u}(v_m), val_u(v_m))}{t}$$

It is reasonable, though not mandatory, to assume that l is the quadratic loss function $((P_{i,u}(v_m) - val_u(v_m))^2)$, because according to a well-known result of (Brier 1950) the quadratic loss function maximizes the forecaster's expected success if she identifies her predictions with her credences (in the literature such scoring rules are also called *proper* scoring rules; a relevant alternative in this respect is, e.g., the negative logarithmic scoring rule taking the negative of the logarithm of the predicted value which proved to be true).

The decisive difference of this setting compared to the previous one is that now the success rates of the candidate methods and the meta-inductive algorithm are relative to [724] elements of the value space: Each method has a success rate for each value v_m. Based on this we can define a weight $w_{i,t}(v_m)$ of method i for predicting event value v_m up to time t as follows (where $s_{aggr,t}$ is the per-round success rate of P_{aggr^*} as defined below):

$$w_{i,t}(v_m) = \frac{max(0, s_{i,t}(v_m) - s_{aggr^*,t}(v_m))}{\sum_{j=1}^{n} max(0, s_{j,t}(v_m) - s_{aggr^*,t}(v_m))}$$

Finally, based on these weights we might define a probabilistic aggregating algorithm as follows:

$$P_{aggr^*,t+1}(v_m) = \sum_{i=1}^{n} w_{i,t}(v_m) \cdot P_{i,t+1}(v_m)$$

It is easy to see that the no regret optimality result of the foregoing section holds for such a meta-level method for each value v_m of E's value space: The probabilistic aggregating forecasting algorithm P_{aggr^*} will approximate the maximum of the success rates of the best candidate methods accessible in the setting regarding each v_m. However, there is a problem: It can easily happen that the methods which are best at a given round depend on the value of the value space. In other words, the meta-inductive forecaster uses weights resulting from different prediction games which can lead to the result that its aggregated probabilities are *incoherent*. To see this, consider the following example:

- Let E be a series of discrete random variables $E_1, E_2, \ldots$.
- $k = 3$, i.e. the value space consists of v_1, v_2, v_3.
- Let $n = 2$, i.e. the accessible candidate methods are P_1 and P_2. Now, let up to round u candidate method P_1 be a perfect expert in predicting v_1 and P_2 be a perfect expert in predicting v_2. Let up to round u P_1 completely fail regarding the predictions of v_2, v_3 and P_2 completely fail regarding predictions of v_1, v_3. Thus for all $t \leq u$: if $val_t(v_1) = 1$, then $P_{1,t}(v_1) = 1$ and $P_{2,t}(v_1) = 0$; and if $val_t(v_2) = 1$, then $P_{2,t}(v_2) = 1$ and $P_{1,t}(v_2) = 0$. Moreover if $val_t(v_3) = 1$ both fail, i.e. $P_{1,t}(v_3) = P_{2,t}(v_3) = 0$.
- So, the candidate predictions are such that their success rates at each round $t \leq u$ (for all convex loss functions without an additive term) are:

	$s_{1,t}(v_i)$	$s_{2,t}(v_i)$
v_1	100%	0%
v_2	0%	100%
v_3	0%	0%

- But then $w_{1,u+1}(v_1) = 1$, thus $P_{aggr^*,u+1}(v_1) = P_{1,u+1}(v_1)$ and $w_{2,u+1}(v_2) = 1$, thus $P_{aggr^*,u+1}(v_2) = P_{2,u+1}(v_2)$. Now assume that at round $u + 1$ both of the candidate methods predict the value they were absolute experts up to round u, i.e. $P_{1,u+1}(v_1) = 1$ and $P_{2,u+1}(v_2) = 1$. Then the predictions of the algorithm are

  $$P_{aggr^*,u+1}(v_1) = 1 \text{ and } P_{aggr^*,u+1}(v_2) = 1$$

  which is probabilistically inconsistent.

So, although each individual provides a probabilistic forecast, pooling the forecasts according to this idea ends up with a forecast that is no longer probabilistically consistent. Regarding each value of the value space such a forecast is no regret optimal, however, this optimality comes at cost of consistency.

One can restore consistency by normalising P_{aggr^*}. Here the idea is to still calculate for each candidate method success rates that depend on the method's success regarding a [725] specific value v_m of the value space. These success rates are then, in a second step, used for defining value-dependent weights for each candidate method. And these weights are again, in a third step, used to construct a prediction as above. However, additionally as a fourth step these predictions are normalized in order to guarantee probabilistic consistency. Such a normalized average probability aggregation algorithm can be defined as follows:

$$P_{aggr^{**},t+1}(v_m) = \frac{P_{aggr^{*},t+1}(v_m)}{\sum\limits_{j=1}^{k} P_{aggr^{*},t+1}(v_j)}$$

A schema of such an implementation is illustrated in figure 2 (more details on this figure see below): Probabilistic forecasts consist no longer of parallel prediction games, but of combining parallel predictions by help of normalization to a single probabilistic forecast.

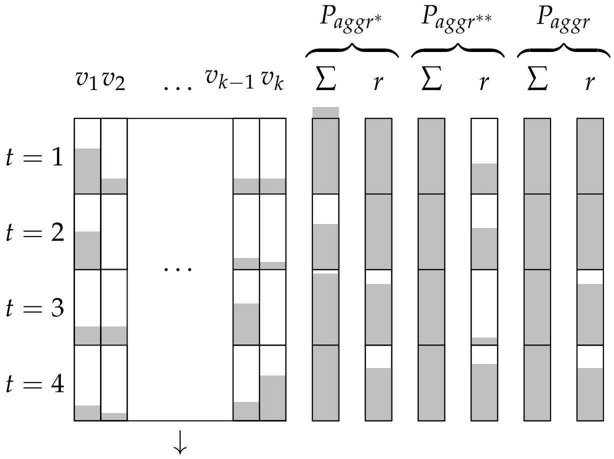

Figure 2: Example of launching k prediction games about single events parallel (P_{aggr^*}), one for each value of the value space. Out of the parallel prediction games a probabilistic forecast about all values is constructed by normalization ($P_{aggr^{**}}$); P_{aggr} constructs its predictions out of averaging the success-rates among the values. The bars under $\sum$ indicate the sum of the meta-inductive algormithm’s probability forecast. The bars under r (regret) indicate proven upper bounds for the regrets. As can be seen, P_{aggr^*}’s regret vanishes in the long run, hower its forecast is probabilistcally inhoherent (does not sum up to 1). $P_{aggr^{**}}$ is probabilistcally coherent through normalization, however, its average per-round regret does not vanish in the long run. And finally, P_{aggr} (cf. the definition below) has advantages of both worlds: it is probabilistically coherent and no regret optimal.

By help of an example one can show that the probabilistic aggregating forecasting algorithm is not optimal with respect to the single values. To see this, consider the following probabilistic prediction game:

- [726] Let us assume that we have three values v_1, v_2, v_3, two forecasters P_1, P_2 and for simplicity reasons let us assume that each of them gives at each round full probability to one of the values. Now, let us assume that the forecasts and the outcome are as follows:

t	1	2	3	4	5	6	7	8	…
P_1	$v_1 : 1.0$	$v_1 : 1.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 1.0$	$v_1 : 1.0$	$v_1 : 0.0$	$v_1 : 0.0$	…
	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	$v_2 : 0.0$	…
	$v_3 : 0.0$	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 1.0$	$v_3 : 0.0$	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 1.0$	…
P_2	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	$v_1 : 0.0$	…
	$v_2 : 1.0$	$v_2 : 0.0$	$v_2 : 1.0$	$v_2 : 0.0$	$v_2 : 1.0$	$v_2 : 0.0$	$v_2 : 1.0$	$v_2 : 0.0$	…
	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 0.0$	$v_3 : 1.0$	$v_3 : 0.0$	$v_3 : 1.0$	…
val	v_1	v_1	v_2	v_3	v_2	v_1	v_2	v_3	…

- Let us furthermore assume a linear loss function (similar counterexamples are possible with other convex loss functions). Then the success rates will converge to $s_{1,t\to\infty}(v_1) = s_{2,t\to\infty}(v_2) = 7/8$, $s_{1,t\to\infty}(v_2) = s_{2,t\to\infty}(v_1) = 5/8$, $s_{1,t\to\infty}(v_3) = s_{2,t\to\infty}(v_3) = 7/8$. Thus, after some point in time t^*, P_1 will gain full attractivity and weight in predicting v_1, P_2 full attractivity and weight in predicting v_2, and both get equal weight in predicting v_3. Hence, starting at $t^* + 1$ the unnormalized and the normalized predictions of the meta-level agents are:

t	t^*	$t^* + 1$	$t^* + 2$	$t^* + 3$	…
P_{aggr^*}	$v_1 : 1.0$	$v_1 : 1.0$	$v_1 : 0.0$	$v_1 : 0.0$	…
	$v_2 : 1.0$	$v_2 : 0.0$	$v_2 : 1.0$	$v_2 : 0.0$	…
	$v_3 : 0.0$	$v_3 : 0.5$	$v_3 : 0.5$	$v_3 : 1.0$	…
$P_{aggr^{**}}$	$v_1 : 0.5$	$v_1 : 0.\overline{66}$	$v_1 : 0.0$	$v_1 : 0.0$	…
	$v_2 : 0.5$	$v_2 : 0.0$	$v_2 : 0.\overline{66}$	$v_2 : 0.0$	…
	$v_3 : 0.0$	$v_3 : 0.\overline{33}$	$v_3 : 0.\overline{33}$	$v_3 : 1.0$	…
val	v_1/v_2	v_1	v_2	v_3	…

- But then—given, e.g., the natural loss function—the success rates of $P_{aggr^{**},t\to\infty}$ are: $s_{aggr^{**},t\to\infty}(v_1) = 19/24 < 7/8 = s_{1,t\to\infty}(v_1)$, $s_{aggr^{**},t\to\infty}(v_2) = 19/24 < 7/8 = s_{2,t\to\infty}(v_2)$, and $s_{aggr^{**},t\to\infty}(v_3) = 10/12 < 7/8 = s_{1,t\to\infty}(v_3) = s_{2,t\to\infty}(v_3)$.

- Hence, regarding all three values $P_{aggr^{**}}$ is no regret *sub*optimal.

As the examples above show, one cannot have both, consistency and optimality with respect to each value of the value space. However, we can construct a probabilistic aggregation method, call it P_{aggr}, that is both coherent and no regret optimal. We can do so simply by averaging the success-rates for the individual values of the value space.

To recognize this, we hint to the mathematical fact that if the loss function l is convex with respect to all values of the value space, then also averaging among the losses with respect to all values of the value space and with respect to all points in time up to the round [727] of consideration is convex (for details see the proof of theorem 2 in the appendix). Let us first define such an average loss measure l_{av}:

$$l^{av}_{i,t} = \frac{\sum\limits_{u=1}^{t} \sum\limits_{m=1}^{k} l(P_{i,u}(v_m), val_u(v_m))}{t \cdot k}$$

We can then define a measure for average success s^{av} which is not relativized to a single value of the value space:

$$s^{av}_{i,t} = 1 - l^{av}_{i,t}$$

Based on these per-round average success rate we can define average success-based weights w^{av} that are also not relativized to a single value of the value space. In order to avoid the need of *cutting off* as described in figure 1, we will simply put the cumulative success together with a learning parameter η in the exponent (for more details on η cf. the proof of theorem 2 in the appendix; exponential weighting by help of a learning parameter is a general strategy to avoid the need of *cutting off* (Cesa-Bianchi and Lugosi 2006, chpt.2)):

$$w^{av}_{i,t} = \frac{e^{\eta \cdot s^{av}_{i,t} \cdot t}}{\sum\limits_{j=1}^{n} e^{\eta \cdot s^{av}_{j,t} \cdot t}}$$

Now, we can define the meta-inductive algorithm for weighted average probability aggregation (AGGR) based on these weights in accordance with (MI):

$$P_{aggr,t+1}(v_j) = \sum_{i=1}^{n} w^{av}_{i,t} \cdot P_{i,t+1}(v_j) \quad \forall j \in \{1, \ldots, v_k\} \tag{AGGR}$$

Since (AGGR) is an instance of (MI) and since l used to determine the weights w^{av} is convex, we can transfer the no regret optimality result of P_{mi} to P_{aggr}:

Theorem 2. *Given the loss function l is convex Pr_{aggr} is a no regret algorithm for aggregating probabilities:*

$$\lim_{t \to \infty} s^{av}_{i,t} - s^{av}_{aggr,t} \;\; \leq \;\; 0 \;\; \forall i \in \{1, \ldots, n\}$$

(For a proof of theorem 2 see the appendix.)

To illustrate this fact we can come back to the last example on the *sub*optimality of $P_{aggr^{**}}$ regarding each value of the value space: Here it was the case that the candidate method P_1 was better than $P_{aggr^{**}}$ regarding v_1 and v_3, P_2 was better than $P_{aggr^{**}}$ regarding v_2 and v_3. However, as calculating the average success-rates demonstrates as an instance of our general result above, both of them are not better than P_{aggr} in averaging over their per-round success-rates regarding all values of the value space v_1, v_2, v_3: $s^{av}_{aggr,t\to\infty} \geq s^{av}_{1,t\to\infty}$ and $s^{av}_{aggr,t\to\infty} \geq s^{av}_{2,t\to\infty}$.

5 Meta-Induction Based on Generalized Brier Scoring

In section 2 we have outlined that the general constraints of *universality* (U), *zero unanimity* (Z), and *irrelevance of alternatives* (I) characterize linear probability aggregation (LIN). In section 3 we referred to the fact that meta-inductive predictions of single values (MI), i.e. deterministic predictions, are optimal in the sense that they provide no regret predictions in [728] the long run. In section 4 we have shown how the optimality of deterministic predictions can be transferred to probabilistic predictions (AGGR). We have done so by defining a measure for the average loss and per-round success. Note that whereas (LIN) was underdetermined regarding the choice of weights, (AGGR) is already much more determined in the sense that for satisfying the constraint of long run optimality (i.e. having no regret in the long run) a certain kind of success-based weighting is sufficient. In this section we want to briefly relate our method of success-based probability aggregation to approaches that link weights to scoring.

Scoring rules are intended to answer the question of measuring the accuracy of probabilistic predictions. Seminal regarding this task became (Brier 1950) where out of practical urgency of evaluating meteorological forecasts a specific score was defined, the so-called *Brier score*. Glenn Brier suggested to *verify* a forecasting method P_i via low scores calculated as (Brier 1950, p.1, equ.2):

$$\frac{1}{t}\sum_{u=1}^{t}\sum_{m=1}^{k}(P_{i,u}(v_m) - val_u(v_m))^2$$

where $val_u(v_m) = 1$, if v_m occurred at time u, and $val_u(v_m) = 0$, if v_m did not occur at time u. A perfect probabilistic predictor up to time t predicts all event outcomes with probability 1 and all the other values with probability 0, i.e. $P_{i,u}(v_m) = val_u(v_m)$ (for all $u \leq t$, $v_m \in \{v_1, \ldots, v_k\}$), and hence has score 0. In the worst case a predictor predicts up to time t a value not showing up with probability 1 and the true event outcome as well as all the other values with probability 0. Then at each round it scores by $2/t$ (for the predicted value $1/t$ and for the prediction of the true outcome $1/t$—for all the other predicted values 0), hence its score up to round t is $t \cdot 2/t = 2$, so the interval of the Brier score is $[0, 2]$. In order to normalize the Brier score, we need to divide it by 2. So, the *Brierian* average loss of prediction method P_i up to round t is:

$$l_{i,t}^{Brier} = \frac{1}{2t} \sum_{u=1}^{t} \sum_{m=1}^{k} (P_{i,u}(v_m) - val_u(v_m))^2$$

Now, recall that we defined the average per-round loss of some prediction method P_i up to round t which underlies the weights of (AGGR) as:

$$l_{i,t}^{av} = \frac{1}{kt} \sum_{u=1}^{t} \sum_{m=1}^{k} l(P_{i,u}(v_m), val_u(v_m))$$

Note that if we define l as a quadratic loss function with some factor, namely as $l(x, y) = \frac{k}{2} \cdot (x - y)^2$, then l^{av} is identical to the normalized version of the Brier score as defined in (Brier 1950): $l_{i,t}^{av} = l_{i,t}^{Brier}$. Since the quadratic loss function is convex (and also the quadratic loss function with a constant factor), also the no regret optimality result holds for linear aggregation where the weights are determined via the normalized Brier score l^{Brier}.

Meta-inductive probability aggregation allows to prove optimality for any convex loss function. Since the normalized Brier score is a specific convex loss function, we can [729] straightforwardly implement it to meta-inductive probability aggregation as defined in (AGGR). This allows for *optimal* probability aggregation based on generalized Brier scoring.

6 Conclusion

In this paper we have argued for a new solution to the problem of weighted probability aggregation. We have seen that some general constraints determine families of aggregation rules like linear aggregation rules. In order to address the problem of specifying such rules by fixing weights we have argued for a success-based calculation of weights as is suggested also in the literature on scoring probabilistic forecasts (Genest and McConway 1990, pp.57ff). As we have shown, such an approach can be justified by help of results on predictions under expert advice since a success-based calculation of weights allows for no regret optimal probabilistic aggregation. This form of meta-inductive probability aggregation can be considered as a generalization of approaching probability aggregation by help of scoring.

Appendix

Here we provide a proof of theorem 2 which is a slight expansion of a proof provided in (Feldbacher-Escamilla 2020), which itself is loosely based on a proof provided in (Shalev-Shwartz and Ben-David 2014, p.253): The main strategy of the proof is to apply inequalities such that the differences of the success rates are narrowly bounded. As we demonstrate now, success-based weighting allows for an optimal bound in the sense that in the limit such weighting cannot be outperformed by any other inference method in terms of the success rate.

Proof. In order to prove the no regret-property of the aggregating method P_{aggr}, we characterise the difference between the competing predictors and that of the aggregating predictor by help of a learning parameter η which is a function of the number of rounds t, and which grows sublinearly with t. If such a characterisation succeeds, then the difference of the success rate grows sublinearly only and vanishes in the limit; this means that by help of such a characterisation the aggregating predictor is shown to be not outperformed by any other predictor in the limit. As it turns out, one can characterise such differences in successes by help of choosing $\eta = \sqrt{\frac{2 \cdot \ln(n)}{T}}$. Here T is an arbitrary round and sometimes also called the *prediction horizon* up to which a boundary is proven (Cesa-Bianchi and Lugosi 2006, p.15). In order to generalise this boundary to any round t, one needs, in a second step, to get rid of the exact choice of T by employing the so-called *doubling trick*, according to which for each round t it is assumed that the prediction horizon T doubles; this assumption increases the bound a bit, but does not change anything regarding the limiting case, and hence allows for proving a general optimality result too. In the following proof we demonstrate the first part (for arbitrary T); the second part of applying the doubling trick can be recapitulated by help of (Mohri, Rostamizadeh, and Talwalkar 2012, p.158).

i. Recall from section 3 and 4 that the probabilistic aggregation method we are aiming at is defined as the weighted ($w_{i,t}$) average of the individual predictions ($P_{i,t}$), where the weights are a function of the per round successes $s_{i,t}$ and the latter are just defined as the "inverse" (within the unit interval) of the losses $l(P_{i,t}(v), val_t(v))$.

ii. Let $\eta = \sqrt{\frac{2 \cdot \ln(n)}{T}}$. Furthermore let l be convex. Let us also restate the weights $w^{av}_{i,t}$ recursively via defining coefficients c: Let $c_{i,1}$ (for $1 \leq i \leq n$) be 1. [730] Then define recursively $c_{i,t+1} = c_{i,t} \cdot e^{-\eta \cdot \sum_{m=1}^{k} l^m_{i,t}/k}$, where $l^m_{i,t} = l(P_{i,t}(v_m), val(v_m))$ is the loss of i at round t with respect to the prediction of all value v_m.

iii. By definition of c we get the following equalities about the ratio of the denominators used in normalisation of the weights (the normalising denominator for $t+1$ and that of t):

$$\frac{\sum\limits_{i=1}^{n} c_{i,t+1}}{\sum\limits_{j=1}^{n} c_{j,t}} = \sum_{i=1}^{n} \frac{c_{i,t+1}}{\sum\limits_{j=1}^{n} c_{j,t}} = \sum_{i=1}^{n} \frac{c_{i,t} \cdot e^{-\eta \cdot \sum\limits_{m=1}^{k} l^m_{i,t}/k}}{\sum\limits_{j=1}^{n} c_{j,t}}$$

$$= \sum_{i=1}^{n} w^{av}_{i,t} \cdot e^{-\eta \cdot \sum\limits_{m=1}^{k} l^m_{i,t}/k}$$

In what follows we abbreviate $\sum\limits_{m=1}^{k} l^m_{i,t}/k$ simply by $\Sigma l_{i,t}$.

iv. [731] By the inequality $e^{-x} \leq 1 - x + \frac{x^2}{2}$ (valid for all $x \geq 0$) we get the instance:

$$e^{-\eta \cdot \Sigma l_{i,t}} \quad \leq \quad 1 - \eta \cdot \Sigma l_{i,t} + \frac{\eta^2 \cdot \left(\Sigma l_{i,t}\right)^2}{2}$$

Note that due to the assumptions in ii. $0 \leq \eta < 1$ and due to the boundedness of loss l by $[0,1]$ $\eta \cdot \Sigma l_{i,t} \in [0,1]$.

v. By substituting the right term in the inequality of iv. for the e-term in iii. we get:

$$\frac{\sum\limits_{i=1}^{n} c_{i,t+1}}{\sum\limits_{j=1}^{n} c_{j,t}} \quad \leq \quad \sum_{i=1}^{n} w_{i,t}^{av} \cdot \left(1 - \eta \cdot \Sigma l_{i,t} + \frac{\eta^2 \cdot \left(\Sigma l_{i,t}\right)^2}{2}\right)$$

and by arithmetic transformation:

$$\leq \quad \sum_{i=1}^{n} w_{i,t}^{av} \; - \left(\eta \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t}\right) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \left(\Sigma l_{i,t}\right)^2\right)\right)$$

By the normalisation of w: $\sum\limits_{i=1}^{n} w_{i,t}^{av} = 1$, so:

$$\leq \quad 1 \; - \left(\eta \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t}\right) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \left(\Sigma l_{i,t}\right)^2\right)\right)$$

By taking the ln on both sides of the inequality:

$$\ln\left(\frac{\sum\limits_{i=1}^{n} c_{i,t+1}}{\sum\limits_{j=1}^{n} c_{j,t}}\right) \quad \leq \quad \ln\left(1 - \left(\eta \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t}\right) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \left(\Sigma l_{i,t}\right)^2\right)\right)\right)$$

vi. By the inequality $e^{-x} \geq 1 - x$ (valid for any x) we get $\ln(e^{-x}) \geq \ln(1-x)$ and hence $-x \geq \ln(1-x)$. So, as an instance:

$$-\left(\eta \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t}\right) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t}^2\right)\right) \quad \geq$$

$$\ln\left(1 - \left(\eta \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t}\right) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t}^2\right)\right)\right)$$

Verify that due to the assumptions in ii. $0 \leq \eta < 1$, the boundedness of loss l by $[0,1]$, as well as the normalisation of w our instance of x is within $[0,1]$.

vii. By substituting the left (upper) term in the inequality of vi. for the right term in the inequality in v. we get:

$$\ln\left(\frac{\sum\limits_{i=1}^{n} c_{i,t+1}}{\sum\limits_{j=1}^{n} c_{j,t}}\right) \;\leq\; -\left(\eta \cdot \sum_{i=1}^{n}\left(w_{i,t}^{av} \cdot \Sigma l_{i,t}\right) - \frac{\eta^2}{2} \cdot \sum_{i=1}^{n}\left(w_{i,t}^{av} \cdot \left(\Sigma l_{i,t}\right)^2\right)\right)$$

and by arithmetic transformation:

$$\leq\; \frac{\eta^2}{2} \cdot \underbrace{\sum_{i=1}^{n}\left(w_{i,t}^{av} \cdot \left(\Sigma l_{i,t}\right)^2\right)}_{\leq 1} - \eta \cdot \sum_{i=1}^{n}\left(w_{i,t}^{av} \cdot \Sigma l_{i,t}\right)$$

$\ldots$ due to $\sum\limits_{i=1}^{n} w_{i,t}^{av} = 1$, and $l \in [0,1]$, so:

$$\leq\; \frac{\eta^2}{2} \cdot\; 1 \;- \eta \cdot \sum_{i=1}^{n}\left(w_{i,t}^{av} \cdot \Sigma l_{i,t}\right)$$

viii. So, we arrived at the inequality (from vii.):

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

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

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

So, we arrive at:

$$\ln\left(\sum_{i=1}^{n} c_{i,T+1}\right) - \ln\underbrace{\left(\sum_{i=1}^{n} c_{i,1}\right)}_{=n} \;\leq\; \frac{T \cdot \eta^2}{2} - \eta \cdot \sum_{t=1}^{T} \sum_{i=1}^{n}\left(w_{i,t}^{av} \cdot \Sigma l_{i,t}\right)$$

Hence:

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

[732] Recall, $c_{i,t}$ is the cumulative loss up to t in the exponent and we are after the bound for the regret with respect to the best predictor, hence we concentrate on the predictor with minimal cumulative loss up to T: Let us denote this predictor with b ($b = (\iota i)(\sum_{t=1}^{T} \Sigma l_{i,t} = min(\sum_{t=1}^{T} \sum l_{1,t}, \ldots, \sum_{t=1}^{T} \sum l_{n,t})))$. If there are more, then we can randomly pick one. Now:

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

Hence:

$$\ln(c_{b,T}) - \ln(n) \quad \leq \quad \frac{T \cdot \eta^2}{2} - \eta \cdot \sum_{t=1}^{T} \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t} \right)$$

ix. By definition of c:

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

So:

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

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

$$-\eta \cdot \sum_{t=1}^{T} \Sigma l_{b,t} - \ln(n) \quad \leq \quad \frac{T \cdot \eta^2}{2} - \eta \cdot \sum_{t=1}^{T} \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t} \right)$$

And by arithmetical transformation:

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t} \right) - \sum_{t=1}^{T} \Sigma l_{b,t} \quad \leq \quad \frac{T \cdot \eta}{2} + \frac{\ln(n)}{\eta}$$

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

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot \Sigma l_{i,t} \right) \; - \sum_{t=1}^{T} \Sigma l_{b,t} \quad \leq \quad \sqrt{2 \cdot \ln(n) \cdot T}$$

Now, what is left is to employ the left term of the difference in the inequality above for proving a bound for the meta-inductive method's regret.

x. According to AGGR, P_{aggr} predicts as follows: $P_{aggr,t}(v_m) = \sum_{i=1}^{n} w_{i,t}^{av} \cdot P_{i,t}(v_m)$. Hence its loss for value m is: $l \left(\sum_{i=1}^{n} (w_{i,t}^{av} \cdot P_{i,t}(v_m)), val_t(v_m) \right)$. And hence its average cumulative loss is:

$$\sum_{t=1}^{T} \sum_{m=1}^{k} l \left(\sum_{i=1}^{n} (w_{i,t}^{av} \cdot P_{i,t}(v_m)), val_t(v_m) \right) / k$$

[733] Since l is convex (according to ii.), we get:

$$\sum_{m=1}^{k} l \left(\sum_{i=1}^{n} (w_{i,t}^{av} \cdot P_{i,t}(v_m)), val_t(v_m) \right) / k \;\; \leq \;\; \sum_{m=1}^{k} \sum_{i=1}^{n} \left(w_{i,t}^{av} \cdot l(P_{i,t}(v_m), val_t(v_m)) \right) / k$$

(I.e.: The loss of a weighted average of predictions is smaller than or equal to the weighted average of the losses of the predictions.) Hence, from the last inequality in ix. and the convexity of l we get:

$$\underbrace{\sum_{t=1}^{T}\sum_{m=1}^{k}\left(l\left(\sum_{i=1}^{n}(w_{i,t}^{av}\cdot P_{i,t}(v_m)), val_t(v_m)\right)\right)/k-\sum_{t=1}^{T}\sum_{m=1}^{k}l(P_{b,t}(v_m), val_t(v_m))/k}_{=l^{av}_{aggr,T}\cdot T-l^{av}_{b,T}\cdot T}$$

$$\leq \quad \sqrt{2\cdot\ln(n)\cdot T}$$

xi. Now, since $s^{av}_{i,T} = 1 - l^{av}_{i,T}$, this means that:

$$s^{av}_{b,T} - s^{av}_{aggr,T} \leq \frac{const}{\sqrt{T}}$$

By applying the above mentioned *doubling trick*, this holds for all T, hence:

$$\lim_{t\to\infty} s^{av}_{b,t} - s^{av}_{aggr,t} \leq 0$$

Since P_b was the method with least cumulative loss up to t (we defined b this way in viii.), this bound holds also with respect to all other predictors (for all $1 \leq i \leq n$).

□

[734]

References

Arrow, Kenneth Joseph (1963). *Social Choice and Individual Values*. 2nd Edition. Yale: Yale University Press.

Brier, Glenn W. (1950). “Verification of Forecasts Expressed in Terms of Probability”. In: *Monthly Weather Review* 78.1, pp. 1–3.

Cesa-Bianchi, Nicolo and Lugosi, Gabor (2006). *Prediction, Learning, and Games*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511546921.

Dietrich, Franz, Endriss, Ulle, Grossi, Davide, Pigozzi, Gabriella, and Slavkovik, Marija (2014). “JA4AI – Judgment Aggregation for Artificial Intelligence (Dagstuhl Seminar 14202)”. In: *Dagstuhl Reports* 4.5. Ed. by Dietrich, Franz, Endriss, Ulle, Grossi, Davide, Pigozzi, Gabriella, and Slavkovik, Marija, pp. 27–39. DOI: 10.4230/DagRep.4.5.27.

Feldbacher-Escamilla, Christian J. (2020). “An Optimality-Argument for Equal Weighting”. In: *Synthese* 197.4, pp. 1543–1563. DOI: 10.1007/s11229-018-02028-1.

Genest, Christian and McConway, Kevin J. (1990). “Allocating the Weights in the Linear Opinion Pool”. In: *Journal of Forecasting* 9.1, pp. 53–73. DOI: 10.1002/for.3980090106.

Genest, Christian, McConway, Kevin J., and Schervish, Mark J. (1986-06). “Characterization of Externally Bayesian Pooling Operators”. In: *The Annals of Statistics* 14.2, pp. 487–501. DOI: 10.1214/aos/1176349934.

Genest, Christian and Zidek, James V. (1986-02). “Combining Probability Distributions: A Critique and an Annotated Bibliography”. In: *Statistical Sciences* 1.1, pp. 114–135.

Grossi, Davide and Pigozzi, Gabriella (2014). *Judgment Aggregation: A Primer*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Williston: Morgan & Claypool.

Kornhauser, Lewis A. and Sager, Lawrence G. (1986). “Unpacking the Court”. In: *The Yale Law Journal* 96.1, pp. 82–117. URL: http://www.jstor.org/stable/796436.

Lehrer, Keith and Wagner, Carl (1981). *Rational Consesus in Science and Society. A Philosophical and Mathematical Study*. Dordrecht: Reidel Publishing Company.

List, Christian and Pettit, Philip (2002). “Aggregating Sets of Judgments: An Impossibility Result”. In: *Economics and Philosophy* 18.01, pp. 89–110.

Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet (2012). *Foundations of Machine Learning*. Cambridge, Massachusetts: The MIT Press.

Rossi, Francesca, Venable, Kristen Brent, and Walsh, Toby (2011). *A Short Introduction to Preferences. Between Artificial Intelligence and Social Choice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Williston: Morgan & Claypool.

Schurz, Gerhard (2008). “The Meta-Inductivist’s Winning Strategy in the Prediction Game: A New Approach to Hume’s Problem”. In: *Philosophy of Science* 75.3, pp. 278–305. DOI: 10.1086/592550.

Shalev-Shwartz, Shai and Ben-David, Shai (2014). *Understanding Machine Learning. From Theory to Algorithms*. Cambridge: Cambridge University Press.