"Calibeating": Beating Forecasters at Their Own Game*

Dean P. Foster† Sergiu Hart‡

May 30, 2022

Abstract

In order to identify expertise, forecasters should not be tested by their calibration score, which can always be made arbitrarily small, but rather by their Brier score. The Brier score is the sum of the calibration score and the refinement score; the latter measures how good the sorting into bins with the same forecast is, and thus attests to "expertise." This raises the question of whether one can gain calibration without losing expertise, which we refer to as "calibeating." We provide an easy way to calibeat any forecast, by a deterministic online procedure. We moreover show that calibeating can be achieved by a stochastic procedure that is itself calibrated, and then extend the results to simultaneously calibeating multiple procedures, and to deterministic procedures that are continuously calibrated.

Contents

*Previous versions: February 2020; October 2021 (Hebrew University of Jerusalem, Center for Rationality DP-743). We thank Drew Fudenberg and Benjy Weiss for useful comments. A presentation is available at `http://www.ma.huji.ac.il/hart/pres.html#calib-beat-p`

†Department of Statistics, Wharton, University of Pennsylvania, Philadelphia, and Amazon, New York. *e-mail*: `dean@foster.net` *web page*: `http://deanfoster.net`

‡Institute of Mathematics, Department of Economics, and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem. *e-mail*: `hart@huji.ac.il` *web page*: `http://www.ma.huji.ac.il/hart`

1 Introduction

Forecasters—whether of weather or of events like elections and sports—make probabilistic predictions, such as "the probability of rain is p." What does it mean, and how does one test whether it is any good? Taking the classic view of probability as long-run frequency, the above prediction translates to "in the days when the forecast is p the frequency of rain is close to p in the long run." If this holds for all values of p used as forecasts, one says that the forecaster is *calibrated*.[1] More generally, the *calibration score* $\mathcal{K}$ is defined as the average squared distance between forecasts and realized frequencies[2] (where each

[1]There is a large literature on calibration; see the survey of Olszewski (2015), and the recent paper of Foster and Hart (2021), which also discusses the economic utility of calibration (see Section I.A there).

[2]By "frequency" we always mean "relative frequency," i.e., the proportion of, say, rainy days.

forecast p is weighted by how often it has been used); evaluated after t days, this yields[3]

$$\mathcal{K} = \frac{1}{t}\sum_{s=1}^{t}\left(c_s - \bar{a}(c_s)\right)^2,$$

where c_s is the forecast at time s and for each p we denote by $\bar{a}(p) \equiv \bar{a}_t(p)$ the frequency of rain[4] in the days from 1 to t where the forecast was p. Being calibrated means that $\mathcal{K}$ is (close to) 0.

A classic and surprising result of Foster and Vohra (1998) is that one can generate forecasts that are *guaranteed* to be calibrated, no matter what the weather will be. This immediately casts some doubt on whether calibration is the appropriate way to test the expertise of forecasters (there is an extensive literature on "experts" that uses calibration tests to check whether they are indeed experts; see, e.g., the book of Cesa-Bianchi and Lugosi 2006 and the survey of Olszewski 2015).[5]

Day	1	2	3	4	5	6	...	$\mathcal{K}$	$\mathcal{R}$	$\mathcal{B}$
Rain	1	0	1	0	1	0				
F1	100%	0%	100%	0%	100%	0%		0	0	0
F2	50%	50%	50%	50%	50%	50%		0	0.25	0.25

Figure 1: Two calibrated forecasts

Take the following simple and well-known example (see Figure 1). Suppose that the weather alternates between rain on odd days and no rain on even days. Consider two rain forecasters: F1 forecasts 100% on odd days and 0% on even days, and F2 forecasts 50% every day. While both forecasts are well calibrated (the calibration score $\mathcal{K}$ of F1 is 0 every day, and that of F2 is 0 on even days and ≈ 0 on odd days[6]), F1 is clearly a much better and more useful forecaster than F2.

The difference between the two forecasts is underscored by appealing to the classic *Brier* (1950) *score* $\mathcal{B}$, which measures how close the forecasts and the realizations are, by the standard mean squared error formula:

$$\mathcal{B} = \frac{1}{t}\sum_{s=1}^{t}\left(c_s - a_s\right)^2,$$

[3]Weighting each forecast by the proportion of days it has been used from 1 to t is the same as giving weight $1/t$ to each day s from 1 to t.

[4]We refer to rain forecasting for concreteness only. The same applies to any binary event, and, in fact, to multidimensional actions and forecasts; see Section 2.

[5]The fact that the calibration score is not the right way to identify experts does *not* imply that calibration should be ignored; on the contrary, calibration is a useful property for forecasts to satisfy (see Section I.A in Foster and Hart 2021).

[6]Specifically, $1/(4t^2)$, which converges to 0 as $t \to \infty$.

where a_s denotes the weather on day s, with $a_s = 1$ standing for rain and $a_s = 0$ for no rain, and c_s is, as above, the forecast on day s. For F1 the Brier score $\mathcal{B}$ is 0 every day (because $c_s = a_s$ for all s), whereas for F2 it is 1/4 every day (because $(0.5-1)^2 = (0.5-0)^2 = 1/4$). The Brier score thus distinguishes well between the two forecasters ($\mathcal{B} = 0$ vs. $\mathcal{B} = 1/4$), while calibration does not ($\mathcal{K} = 0$ for both).

To interpret this difference in the Brier scores, view forecasting as consisting of two separate ingredients. The first one is the "classification" or "sorting" of days into "bins," where all the days with the same forecast p are assigned to the same bin. The second one is the specific value of the forecast p that is used to define each bin, which we refer to as the "label" of the bin. In the above example, F1 sorts the days into two bins, a 100%-bin, which consists of the odd days, and a 0%-bin, which consists of the even days, whereas for F2 there is a single bin, the 50%-bin, which contains all days. Both bins of F1 are homogeneous: there is no variance among the days in the same bin (they are either all "rain," or all "no rain"); by contrast, in the single bin of F2 there is a large variance among the days (half of them are "rain" and half "no rain"). This "within-bin variance" is captured by the *refinement score* $\mathcal{R}$, which is the average squared distance between the weather a_s and the bin-average weather (which is the average frequency of rain on the days from 1 to t that are in the c_s-bin, i.e., on those days when the forecast was the same as on day s), denoted by $\bar{a}(c_s)$:

$$\mathcal{R} = \frac{1}{t}\sum_{s=1}^{t}\left(a_s - \bar{a}(c_s)\right)^2.$$

The Brier score neatly decomposes into the sum of the refinement and the calibration scores,

$$\mathcal{B} = \mathcal{R} + \mathcal{K}$$

(this easily follows from the equality $\mathbb{E}[X^2] = \mathbb{V}ar[X] + \mathbb{E}[X]^2$; see Section 2.1). The refinement score $\mathcal{R}$ yields the average of the within-bin variances, and the calibration score $\mathcal{K}$ the average squared distance between the bin labels and the bin averages. Perfect calibration, i.e., $\mathcal{K} = 0$, says that all the labels are correct: the label of each bin, i.e., the value of the forecast that defines the bin, is equal to the average weather of the bin. In addition, the refinement score $\mathcal{R}$ and the calibration score $\mathcal{K}$ are "orthogonal": changing the labels does not affect $\mathcal{R}$, and changing the distribution of actions within each bin without changing their average does not affect[7] $\mathcal{K}$. Returning to the example, we have $\mathcal{R} = \mathcal{K} = \mathcal{B} = 0$ for all t for F1, and $\mathcal{R} \approx 1/4$, $\mathcal{K} \approx 0$, $\mathcal{B} = 1/4$ for all t for F2.

Thus, our first conclusion is

Conclusion: Experts better be tested by the Brier score and not by calibration

[7] To get perfect classification without calibration in our example, use, for instance, the forecast 75% on odd days and the forecast 25% on even days: $\mathcal{R} = 0$ and $\mathcal{K} = 1/16$ for all t.

alone.

Unlike the calibration score, the Brier score cannot in general be brought down to zero in the long run. Indeed, for an i.i.d. 50% probability of rain, the refinement score $\mathcal{R}$ is close to $(1/2) \cdot (1/2) = 1/4$ for *any* forecasting sequence (because this is the variance of each bin), and thus the Brier score $\mathcal{B}$ is at least 1/4. However, if there are certain "regularities" or "patterns" in the weather, then an expert forecaster who recognizes them can get a lower refinement score. For example, suppose that it is very likely that when it rains, it does so for precisely two consecutive days; this means a high probability, say 90%, that 1 comes after 01 and also that 0 comes after 011 (where 1 stands for rain and 0 for no rain). For a forecaster that forecasts p_1 if and only if the last two days were 01, and forecasts p_2 (different from p_1) if and only if the last three days were 011, the p_1-bin and the p_2-bin each have a low variance of $0.9 \cdot 0.1 = 0.09$. Knowledge about the weather, which we refer to as *expertise*, is thus reflected in sorting the days into bins that consist of similar days, and in making the binning as refined as possible[8]—that is, in having a low refinement score $\mathcal{R}$.

Returning to calibration, a forecaster can always guarantee its forecasts to be calibrated, by the Foster and Vohra (1998) result.[9] However, this would require it to run one of the calibration procedures (some of which—like the "forecast-hedging" one of Section 5 of Foster and Hart 2021—are extremely simple) and ignore whatever expert knowledge he has about the weather, and whichever patterns he has identified in the data.

Thus, the natural question that arises is

> *Question: Can one gain calibration without losing expertise?*

In formal terms, can one decrease $\mathcal{K}$ to zero without increasing $\mathcal{R}$?

This can of course be always done *in retrospect*: replacing each forecast p with the corresponding bin average $\bar{a}(p)$ yields calibration while preserving the binning, and thus the refinement score $\mathcal{R}$. For example, if the frequency of rain on the days when the forecast was 70% turned out to be 40%, then each forecast of 70% is "corrected" to 40%. The new calibration score is then zero, i.e., $\mathcal{K}' = 0$, while the refinement score is unchanged, i.e., $\mathcal{R}' = \mathcal{R}$; therefore, the Brier score is decreased by the calibration score: $\mathcal{B}' = \mathcal{B} - \mathcal{K}$ (because $\mathcal{B}' = \mathcal{K}' + \mathcal{R}' = 0 + \mathcal{R} = \mathcal{R}$ and $\mathcal{R} = \mathcal{B} - \mathcal{K}$). We will call this

> *"Calibeating": Beating the Brier score by an amount equal to the calibration score.*

The calibeating described above is however obtained only in retrospect—*offline*—since the bin averages are known only at the time t when the testing is done.[10]

[8] Refining the binning can only decrease $\mathcal{R}$; see Section 5 and Appendix A.4.

[9] Such a forecaster is an expert *not* on weather, but on calibration.

[10] Moreover, the forecast corrections depend on the testing horizon t, since the average frequency of rain may well change over time: $\bar{a}_t(p)$ and $\bar{a}_{t'}(p)$ may be quite different for $t \neq t'$.

The interesting question is then what can be done *online*, by a procedure where the forecast of each day s may be modified on the basis of what is known at that time only and nothing beyond it (i.e., neither the upcoming weather on day s, nor the future weather and forecasts on days after s). Our main result is

Result: One can guarantee online calibeating of forecasts.

The first result (Theorem 3 in Section 4) shows that this can be achieved by a simple online procedure: replace each forecast by the average frequency of rain on the *previous* days in which this forecast was made. This attains—*online*—the same lowering of the Brier score by the calibration score that is obtained by the above offline correction. We emphasize that this calibeating is achieved for weather and forecasts to be beaten that are arbitrary (and not stationary in any way), for sorting into bins that may be far from perfect, and for bin averages that need not converge; moreover, everything is guaranteed uniformly, even against a so-called "adversary." The proof uses a neat online estimation of the variance.

Thus, any forecast that is not calibrated can be beaten, online, by another forecast with a strictly better (i.e., lower) Brier score. An alternative interpretation of the result takes a forecasting procedure and announces every period, instead of the intended forecast, its corresponding calibeating replacement (as described in the previous paragraph). This generates a new forecasting procedure, whose Brier score is lower than that of the original one—a clear improvement.[11]

Now the calibeating procedure of our first result need not be calibrated itself, which means that it may be calibeaten too. To avoid this, our second result (Theorem 4 in Section 6) provides a calibeating procedure that is guaranteed to be calibrated, by appealing to a "stochastic fixed point" result, namely, the stochastic "outgoing" tool of Foster and Hart (2021). The calibeating in this case thus yields $\mathcal{K}' = 0$ and $\mathcal{B}' = \mathcal{R}' \leq \mathcal{R}$.

The procedure of this second result is *stochastic*, as it must be in order to guarantee calibration (cf. Dawid 1982, Oakes 1985, and Foster and Vohra 1998). However, if the calibration requirement is weakened to *continuous* calibration—a concept introduced in Foster and Hart (2021), which implies smooth and weak calibration as well, and suffices for equilibrium dynamics—then using instead the fixed point outgoing tool of Foster and Hart (2021) we obtain (Theorem 6 in Section 7) *deterministic* calibeating procedures that are continuously calibrated.

Finally, we show that all the above results can be extended to simultaneously calibeating multiple forecasters (Theorem 8 in Section 8).

To summarize the contribution of this paper: we address the frequently-asked question of how to get better forecasts when there is some expertise. We argue that expertise

[11] This may apply, for instance, to "online regression" or "online least-squares" procedures, introduced by Foster (1991)—see also Forster (1999), Vovk (2001), Azoury and Warmuth (2001), and Cesa-Bianchi and Lugosi (2006)—which minimize the Brier score directly, and need not be calibrated in general.

better be tested by the Brier score and not just by calibration, and show how to calibeat forecasts that are not calibrated, i.e., lower their Brier score by at least their calibration score, without losing the expertise embodied in these forecasts.

2 The Setup

Let A be the set of possible outcomes, which we call *actions*, and let C be the set of *forecasts* about these actions. We assume that $C \subset \mathbb{R}^m$ is a nonempty compact convex subset of a Euclidean space, and that $A \subseteq C$. Some examples: (i) $A = \{0, 1\}$, with $a = 1$ standing for "rain" and $a = 0$ for "no rain," and $C = [0, 1]$, with c in C standing for "the chance of rain is c"; (ii) more generally, C is the set of probability distributions $\Delta(A)$ on a finite set A, i.e., a unit simplex (we identify the elements of A with the unit vectors of C); (iii) C is the convex hull $\text{conv}(A)$ of A. Let $\gamma := \text{diam}(C) \equiv \max_{c,c' \in C} \|c - c'\|$ denote the *diameter* of the set C. Let $\delta > 0$; a subset D of C is a *δ-grid* of C if for every $c \in C$ there is $d \in D$ at a distance less than δ from c, i.e., $\|d - c\| < \delta$; a compact set C always has a finite δ-grid (obtained from a finite subcover by open δ-balls).

The time periods are indexed by $t = 1, 2, \ldots$. An *action sequence* is $\mathbf{a} = (a_t)_{t \geq 1}$ with $a_t \in A$ for all t, and we write $\mathbf{a}_t = (a_s)_{1 \leq s \leq t}$ for its first t elements; similarly, a *forecasting sequence* is $\mathbf{c} = (c_t)_{t \geq 1}$ with $c_t \in C$ for all t, and we put $\mathbf{c}_t = (c_s)_{1 \leq s \leq t}$.

2.1 The Calibration, Refinement, and Brier Scores

Fix a time horizon t. For each possible forecast x in C let[12]

$$\begin{aligned} n_t(x) &:= |\{1 \leq s \leq t : c_s = x\}|, \\ \bar{a}_t(x) &:= \frac{1}{n_t(x)} \sum_{1 \leq s \leq t : c_s = x} a_s, \quad \text{and} \\ v_t(x) &:= \frac{1}{n_t(x)} \sum_{1 \leq s \leq t : c_s = x} \|a_s - \bar{a}_t(x)\|^2 \end{aligned}$$

be, respectively, the *number* of times that the forecast x has been used up to time t, and the action *average* and *variance* in those periods.[13]

The *calibration error* $e_t(x)$ of a forecast x is the difference between the action average and x, i.e.,

$$e_t(x) := \bar{a}_t(x) - x,$$

[12]The number of elements of a finite set Z is denoted by $|Z|$.

[13]When x has not been used, i.e., $n_t(x) = 0$, we put for convenience $v_t(x) := 0$ and (see below) $e_t(x) := 0$.

and the *calibration score* is the average square calibration error, namely,[14,15]

$$\mathcal{K}_t := \sum_{x \in C} \left(\frac{n_t(x)}{t} \right) \|e_t(x)\|^2 ;$$

thus, the error of each x is weighted in proportion to the number of times $n_t(x)$ that x has been used (the weights add up to 1 because $\sum_x n_t(x) = t$). Since from 1 to t there are exactly $n_t(x)$ terms with $x = c_s$, this is equivalent to

$$\mathcal{K}_t = \frac{1}{t} \sum_{s=1}^{t} \|e_t(c_s)\|^2 = \frac{1}{t} \sum_{s=1}^{t} \|\bar{a}_t(c_s) - c_s\|^2 .$$

The *refinement score* is the average over all forecasts of the corresponding action variances:

$$\mathcal{R}_t := \sum_{x \in C} \left(\frac{n_t(x)}{t} \right) v_t(x);$$

again, this is equivalently expressed as

$$\mathcal{R}_t = \frac{1}{t} \sum_{s=1}^{t} \|a_s - \bar{a}_t(c_s)\|^2 .$$

Finally, the *Brier* (1950) *score,*

$$\mathcal{B}_t := \frac{1}{t} \sum_{s=1}^{t} \|a_s - c_s\|^2 ,$$

measures how close the forecasts c_s are to the actions a_s by a standard mean of squared error formula.

One may assume for convenience[16] that one assigns to the bins the *differences* $z_s := a_s - c_s$ between actions and forecasts, instead of the actions a_s; this amounts to subtracting the constant x from all the entries in the x-bin, and then $e_t(x)$ and $v_t(x)$ become, respectively, the expectation and variance of the x-bin. The empirical distribution of the differences z_s and of the bin labels c_s yields two ($\mathbb{R}^m$-valued) random variables, which we denote by Z and U, respectively; namely, the pair (Z, U) takes the value $(z_s, c_s) \equiv (a_s - c_s, c_s)$ for $s = 1, ..., t$ with probability $1/t$ each. With this representation

[14]The sum is finite as it goes over all x with $n_t(x) > 0$, i.e., over x in the set $\{c_1, ..., c_t\}$.

[15]We refer to $\mathcal{K}_t$ as the "ℓ_2-calibration score," to distinguish it from K_t (note the different font) that is used in other papers (e.g., Foster and Hart 2021 and Hart 2021), and which is the "ℓ_1-calibration score," i.e., the weighted average of $\|e_t(x)\|$ rather than $\|e_t(x)\|^2$. The two scores are equivalent, since $(K_t)^2 \leq \mathcal{K}_t \leq \gamma K_t$ (the first inequality by Jensen's inequality, the second by $\|e_t(x)\| \leq \gamma$), and so $\mathcal{K}_t \to 0$ if and only if $K_t \to 0$.

[16]See Foster and Hart (2021); this matters also when generalizing to fractional binnings (Section 7).

we have

$$\begin{aligned} e_t(x) &= \mathbb{E}\left[Z|U=x\right], \\ v_t(x) &= \mathbb{V}ar\left[Z|U=x\right], \\ \mathcal{K}_t &= \mathbb{E}\left[\left\|\mathbb{E}\left[Z|U\right]\right\|^2\right], \\ \mathcal{R}_t &= \mathbb{E}\left[\mathbb{V}ar\left[Z|U\right]\right], \text{ and} \\ \mathcal{B}_t &= \mathbb{E}\left[\left\|Z\right\|^2\right] = \mathbb{E}\left[\mathbb{E}[\left\|Z\right\|^2|U]\right]. \end{aligned}$$

Using the identity $\mathbb{E}\left[X^2\right] = \mathbb{V}ar\left[X\right] + \mathbb{E}\left[X\right]^2$ for each one of the m coordinates of $Z|U$, summing the coordinates, and then taking overall expectation yields

$$\mathcal{B}_t = \mathcal{R}_t + \mathcal{K}_t, \tag{1}$$

which is a useful decomposition of the Brier score (see Sanders 1963 and Murphy 1972).

2.2 Calibration

A stochastic *forecasting procedure* σ is a mapping $\sigma : \cup_{t\geq 1}(A^{t-1} \times C^{t-1}) \to \Delta(C)$; i.e., to each history $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1})$ of actions and forecasts before time t the procedure σ assigns a probability distribution $\sigma(\mathbf{a}_{t-1}, \mathbf{c}_{t-1})$ on C, which yields the forecast $c_t \in C$. When these distributions are all pure (i.e., their support is always a single c_t in C), the procedure is *deterministic*.

Let $\varepsilon \geq 0$; a (stochastic) procedure σ is *ε-calibrated* (Foster and Vohra 1998) if[17,18]

$$\overline{\lim_{t\to\infty}}\left(\sup_{\mathbf{a}_t}\mathbb{E}\left[\mathcal{K}_t\right]\right) \leq \varepsilon^2$$

(the expectation $\mathbb{E}$ is taken over the random forecasts of σ).

2.3 The Concept of "Calibeating"

We come now to the central concept of this paper, "calibeating," which stands for "beating by an amount equal to the calibration score": a forecasting sequence $\mathbf{c}$ "calibeats" another forecasting sequence $\mathbf{b}$ if[19] $\mathbf{c}$ beats the Brier score of $\mathbf{b}$ by at least $\mathbf{b}$'s calibration score (i.e., $\mathcal{B}^{\mathbf{c}} \leq \mathcal{B}^{\mathbf{b}} - \mathcal{K}^{\mathbf{b}}$ in the long run). Thus, if $\mathbf{b}$ is not calibrated, and hence its calibration score

[17] The calibration score $\mathcal{K}_t$ depends on the actions and forecasts up to time t, and is thus a function $\mathcal{K}_t \equiv \mathcal{K}_t(\mathbf{a}, \sigma)$ of the action sequence $\mathbf{a}$ and the forecasting procedure σ (in fact, only $\mathbf{a}^t$ and σ^t matter for $\mathcal{K}_t$). The same applies to the other scores throughout the paper.

[18] The reason that we have ε^2 on the right-hand side is that we are dealing here with the square-calibration score; the same applies to calibeating. The definition here implies the standard one that uses K_t instead of $\mathcal{K}_t$ (e.g., Foster and Hart 2021), since, as we have seen in footnote 15, $(K_t)^2 \leq \mathcal{K}_t$.

[19] Fix the action sequence for now.

$\mathcal{K}^{\mathbf{b}}$ is positive, then the Brier score $\mathcal{B}^{\mathbf{c}}$ of $\mathbf{c}$ is not just better (i.e., lower) than the Brier score $\mathcal{B}^{\mathbf{b}}$ of $\mathbf{b}$, but it is strictly better, by at least $\mathcal{K}^{\mathbf{b}}$. The formal definition will require calibeating to be carried out *online*—i.e., having access only to the current forecast of $\mathbf{b}$ (and the history) and nothing beyond that—and also to be *guaranteed*—i.e., to hold no matter what the sequences of actions and forecasts will be.[20]

By way of the uniformity requirement, we consider a given set $B \subseteq C$ of possible forecasts; for instance, B may be a finite set. A forecasting procedure σ all of whose forecasts are in B is called a *B-forecasting procedure* (when $B = C$ we will usually just say a "forecasting procedure"). Let Σ_B denote the set of all B-forecasting procedures σ, i.e., all mappings $\sigma : \cup_{t\geq 1}(A^{t-1} \times B^{t-1}) \to \Delta(B)$. For $\sigma \in \Sigma_B$, let $b_t \in B$ denote the forecast at time[21] t, and put $\mathbf{b}_t = (b_s)_{1\leq s\leq t}$ and $\mathbf{b} = (b_s)_{s\geq 1}$.

Assume that in each period t the forecast b_t is announced *before* we provide our forecast c_t. Thus, (the distribution of) c_t may depend on $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{b}_t)$, i.e., on the history $h_{t-1} = (\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{b}_{t-1})$ before time t together with the current b_t. A $\mathbf{b}$-*based forecasting procedure* ζ is a mapping[22] $\zeta : \cup_{t\geq 1}(A^{t-1} \times C^{t-1} \times B^t) \to \Delta(C)$. We will use superscripts $\mathbf{b}, \mathbf{c}$ on the scores $\mathcal{B}, \mathcal{R}, \mathcal{K}$ to denote the sequence to which they apply, and similarly for action averages; for example, $\bar{a}_t^{\mathbf{b}}(x)$ is the average of the actions in all periods $s \leq t$ where $b_s = x$, and $\bar{a}_t^{\mathbf{c}}(x)$ is the average of the actions in all periods $s \leq t$ where $c_s = x$.

Let $\varepsilon \geq 0$; a $\mathbf{b}$-based procedure ζ is (ε, B)-*calibeating* if its Brier score beats the Brier score of *any* B-forecasting procedure σ (on which it is based) by that procedure's calibration score; formally,

$$\overline{\lim_{t\to\infty}} \left(\sup_{\sigma\in\Sigma_B} \sup_{\mathbf{a}_t\in A^t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \left(\mathcal{B}_t^{\mathbf{b}} - \mathcal{K}_t^{\mathbf{b}}\right)\right] \right) \leq \varepsilon^2, \tag{2}$$

where the expectation $\mathbb{E}$ is over the random forecasts of[23] of σ and ζ; when $\varepsilon = 0$ we call this B-*calibeating*. Thus, calibeating is guaranteed for any sequence $\mathbf{a}$ of actions and any sequence $\mathbf{b}$ of resulting forecasts of σ, *uniformly* over all B-forecasting procedures σ and action sequences $\mathbf{a}$.

Clearly, condition (2) is not affected if one allows the sequences $\mathbf{a}_t$ to be random. Moreover, since *all* sequences $\mathbf{a}_t$ are considered, one may envision an "adversary" that chooses the B-forecasting procedure σ as well as the action sequence $\mathbf{a}_t$, and so the sequences $\mathbf{b}_t$ and $\mathbf{a}_t$ may well be "coordinated." Thus, $\sup_\sigma \sup_{\mathbf{a}_t}$ in (2) is the same as $\sup_{\mathbf{a}_t,\mathbf{b}_t}$, where $\mathbf{b}_t$ ranges over B^t; indeed, the latter supremum can only be larger, as all

[20] Moreover, it should hold uniformly over all these sequences.

[21] In general, b_t is a random variable.

[22] One should not confuse "B-forecasting" with "$\mathbf{b}$-based"; the former refers to the *outputs* of the procedure (all forecasts are in B) whereas the latter refers to the *inputs* of the procedure (the sequence $\mathbf{b}$).

[23] And so there is no expectation $\mathbb{E}$ when these are deterministic procedures.

sequences $\mathbf{b}_t$ are considered there and not just those generated by σ; however, it cannot be strictly larger since all σ that forecast a fixed sequence $\mathbf{b}_t$ (ignoring the history) are included in the former supremum. Thus, a $\mathbf{b}$-based procedure ζ is (ε, B)-calibeating if

$$\overline{\lim_{t\to\infty}} \left(\sup_{\mathbf{a}_t \in A^t, \mathbf{b}_t \in B^t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \left(\mathcal{B}_t^{\mathbf{b}} - \mathcal{K}_t^{\mathbf{b}} \right) \right] \right) \leq \varepsilon^2, \tag{3}$$

where the expectation is now over the randomizations of ζ.

2.3.1 Calibeating for General B

Since $\mathcal{B} - \mathcal{K} = \mathcal{R}$, we can replace $\mathcal{B}_t^{\mathbf{b}} - \mathcal{K}_t^{\mathbf{b}}$ in (2) and (3) with the refinement score $\mathcal{R}_t^{\mathbf{b}}$ of $\mathbf{b}$: calibeating means that $\mathbf{c}$'s Brier score beats $\mathbf{b}$'s refinement score. This allows generalizing the notion of calibeating to sequences $\mathbf{b} = (b_t)_{t\geq 1}$ for which b_t need not be an element of C. The "forecast" may thus be "a nice day," a "red day," a "b-day," or just "b," for some b in an arbitrary set B. What matters for the resulting refinement scores $\mathcal{R}_t^{\mathbf{b}}$ are the bins into which the days are classified and the ensuing bin variances; the specific labels b of the bins do not matter (the labels *do* however matter for the calibration score, which is "orthogonal" to the refinement score). Therefore, we extend our definition to arbitrary sets B: a $\mathbf{b}$-based forecasting procedure is (ε, B)-*calibeating* if

$$\overline{\lim_{t\to\infty}} \left(\sup_{\sigma \in \Sigma_B} \sup_{\mathbf{a}_t \in A^t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}} \right] \right) \leq \varepsilon^2, \tag{4}$$

or, equivalently,

$$\overline{\lim_{t\to\infty}} \left(\sup_{\mathbf{a}_t \in A^t, \mathbf{b}_t \in B^t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}} \right] \right) \leq \varepsilon^2. \tag{5}$$

As we will see below, this natural extension will be useful, for instance, when considering the joint binning generated by several forecasting procedures.

2.4 Minimax (MM) Procedures and Fixed Point (FP) Procedures

In Foster and Hart (2021) we have introduced the concepts of "procedures of *type MM*," which use a finite minimax[24] construct each period, and "procedures of *type FP*," which use a continuous fixed point construct each period. This distinction, of the polynomial vs. non-polynomial variety, is significant and not just a matter of proof technique (indeed, FP-procedures may be used to get to Nash equilibria in non-zero-sum games); see Sections III.D, VI, and VII in Foster and Hart (2021) for details.

[24] Or linear programming.

3 The Online Refinement Score

The main tool that we will use is that $\mathcal{R}_t$, the refinement score at time t, which is the average variance of the bins and can thus be computed only at time t when the averages of all bins are known (i.e., *offline*), can be approximated by a similar score $\widetilde{\mathcal{R}}_t$, which is computed period by period (i.e., *online*).

Specifically, we define the *online refinement score* $\widetilde{\mathcal{R}}_t$ *at time* t by

$$\widetilde{\mathcal{R}}_t := \frac{1}{t}\sum_{s=1}^{t} \|a_s - \bar{a}_{s-1}(c_s)\|^2 ,$$

where for each c in C we take $\bar{a}_0(c)$ to be an arbitrary element of C. Comparing with the refinement score $\mathcal{R}_t = (1/t)\sum_{s=1}^{t} \|a_s - \bar{a}_t(c_s)\|^2$, we see that what $\widetilde{\mathcal{R}}_t$ does is to replace for each $s = 1, ..., t$ the term $\bar{a}_t(c_s)$, the average at time t of the c_s-bin to which a_s is assigned[25] (an average that will be determined only at time t, i.e., offline), by the term $\bar{a}_{s-1}(c_s)$, the past average (i.e., before time s) of that same c_s-bin, which *is* known at time s (i.e., online).

The following proposition bounds the difference between $\widetilde{\mathcal{R}}_t$ and $\mathcal{R}_t$.

Proposition 1 *For any* $t \geq 1$ *and any sequences* $\mathbf{a}_t$ *and* $\mathbf{c}_t$ *we have*

$$\mathcal{R}_t \leq \widetilde{\mathcal{R}}_t \leq \mathcal{R}_t + \gamma^2 \frac{N_t}{t}\left(\ln\left(\frac{t}{N_t}\right) + 1\right), \tag{6}$$

where $N_t := |\{c_s : 1 \leq s \leq t\}|$ *is the number of distinct elements in the sequence*[26] $\mathbf{c}_t = (c_1, ..., c_t)$.

Thus, $\widetilde{\mathcal{R}}_t - \mathcal{R}_t \to 0$ as $t \to \infty$ when $N_t/t \to 0$, i.e., the number of forecasts used up to time t increases at a slower rate than[27] t. When forecasts belong to a *finite* set $D \subset C$, and so $N_t \leq |D|$ and $\ln(t/N_t) \leq \ln t$ for all t, we get

$$0 \leq \widetilde{\mathcal{R}}_t - \mathcal{R}_t \leq |D|\frac{\ln t + 1}{t}. \tag{7}$$

[25] As pointed out in Section 2.1, neither $\mathcal{R}_t$ nor $\widetilde{\mathcal{R}}_t$ is affected if we assign to the c_s-bin the action a_s or the difference $z_s = a_s - c_s$.

[26] That is, the number of distinct forecasts used in the first t periods.

[27] For a simple example where N_t/t does not converge to 0 and the online refinement score $\widetilde{\mathcal{R}}_t$ does not approach the refinement score $\mathcal{R}_t$, take

$$\begin{array}{lcccccccccccc} \mathbf{a}: & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & ... & 0 & 1 & ... \\ \mathbf{c}: & 1 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{3} & \frac{1}{3} & \frac{1}{4} & \frac{1}{4} & ... & \frac{1}{n} & \frac{1}{n} & ... \end{array}.$$

Indeed, for all even periods $t = 2n$ (where $N_t = n$, and so $N_t/t \to 1/2$), we have $\mathcal{R}_t = 1/4$ (since each $(1/i)$-bin contains two elements, $a_{2i-1} = 0$ and $a_{2i} = 1$) and $\widetilde{\mathcal{R}}_t \geq 1/2$ (since $(a_{2i} - \bar{a}_{2i-1}(c_{2i}))^2 = (1-0)^2 = 1$ and $(a_{2i-1} - \bar{a}_{2i-2}(c_{2i-1}))^2 \geq 0$).

Proposition 1 follows from the following online formula for the variance. Let $(x_n)_{n\geq 1}$ be a sequence of vectors in a Euclidean space (or, more generally, in a normed vector space).

Proposition 2 *For every $n \geq 1$ we have*

$$\sum_{i=1}^{n} \|x_i - \bar{x}_n\|^2 = \sum_{i=1}^{n} \left(1 - \frac{1}{i}\right) \|x_i - \bar{x}_{i-1}\|^2 ,$$

where $\bar{x}_m := (1/m)\sum_{i=1}^{m} x_i$ denotes the average of[28] *$x_1, ..., x_m$.*

Proof. Put $s_n := \sum_{i=1}^{n} \|x_i - \bar{x}_n\|^2$; we claim that

$$s_n = s_{n-1} + \left(1 - \frac{1}{n}\right) \|x_n - \bar{x}_{n-1}\|^2 . \tag{8}$$

We provide a short proof:[29] let $n \geq 2$ (when $n = 1$ both sides vanish), and assume that $\bar{x}_{n-1} = 0$ (this is without loss of generality, since subtracting a constant from all the x_i does not affect any of the terms); then $\bar{x}_n = (1/n)x_n$, and so, using $s_n = \sum_{i=1}^{n} ||x_i||^2 - n||\bar{x}_n||^2$, we get

$$s_n - s_{n-1} = \left(\sum_{i=1}^{n} \|x_i\|^2 - n \left\|\frac{1}{n} x_n\right\|^2\right) - \sum_{i=1}^{n-1} \|x_i\|^2 = \|x_n\|^2 - \frac{1}{n} \|x_n\|^2 ,$$

which is $(1 - 1/n) \|x_n\|^2 = (1 - 1/n) \|x_n - \bar{x}_{n-1}\|^2$.

Applying (8) recursively yields the result. □

Let $v_n := (1/n)\sum_{i=1}^{n} \|x_i - \bar{x}_n\|^2$ denote the variance of $x_1, ..., x_n$, and put $\widetilde{v}_n := (1/n)\sum_{i=1}^{n} \|x_i - \bar{x}_{i-1}\|^2$; i.e., $\bar{x}_n$, the final (up to n) average, is replaced for each $i = 1, ..., n$ with $\bar{x}_{i-1}$, the previous (up to $i - 1$) average.[30] We refer to $\widetilde{v}_n$ as the *online variance* of $x_1, ..., x_n$. Proposition 2 gives $\widetilde{v}_n - v_n = (1/n)\sum_{i=1}^{n} (1/i) \|x_i - \bar{x}_{i-1}\|^2$, and so

$$0 \leq \widetilde{v}_n - v_n \leq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{i} \xi^2 \leq \xi^2 \frac{\ln n + 1}{n} \tag{9}$$

where $\xi := \max_{1\leq i,j\leq n} \|x_i - x_j\|$.

[28] The sum on the right-hand side effectively starts from $i = 2$, and so it does not matter how $\bar{x}_0$ is defined.

[29] An alternative proof of (8) uses the formula

$$\mathbb{V}ar(X) = \mathbb{E}\left[\mathbb{V}ar(X|Y)\right] + \mathbb{V}ar(\mathbb{E}\left[X|Y\right]),$$

where $X = x_i$ with probability $1/n$ and Y is the indicator that $i = n$. Formula (8) is known as a "variance update" formula; see, e.g., Welford (1962).

[30] Take $\bar{x}_0$ to be an arbitrary element of the convex hull of the x_i.

Proposition 1 now easily follows.

Proof of Proposition 1. Let $D \equiv D_t := \{c_s : 1 \le s \le t\} \subset C$ be the set of forecasts used up to time t, i.e., the set of nonempty bins. For each $d \in D$ we apply (9) to get

$$0 \le \frac{1}{n_t(d)} \sum_{s \le t : c_s = d} \|a_s - \bar{a}_{s-1}(d)\|^2 - \frac{1}{n_t(d)} \sum_{s \le t : c_s = d} \|a_s - \bar{a}_t(d)\|^2 \le \gamma^2 \frac{\ln n_t(d) + 1}{n_t(d)}.$$

Averaging over d in D_t with the weights $n_t(d)/t$ then yields

$$0 \le \widetilde{\mathcal{R}}_t - \mathcal{R}_t \le \gamma^2 \frac{1}{t} \sum_{d \in D_t} (\ln n_t(d) + 1).$$

Since the function ln is concave and $\sum_{d \in D_t} n_t(d) = t$, the sum $\sum_{d \in D_t} \ln n_t(d)$ is maximal when all the $n_t(d)$ are equal, i.e., when $n_t(d) = t/N_t$ for each $d \in D_t$; this yields the result (6). □

4 A Simple Way to Calibeat

We provide a simple calibeating procedure. The set B is taken for now to be finite.[31]

Theorem 3 *Let B be a finite set, and let ζ be the deterministic $\mathbf{b}$-based forecasting procedure given by*

$$c_t = \bar{a}^{\mathbf{b}}_{t-1}(b_t) \tag{10}$$

for every time[32] *$t \ge 1$. Then ζ is B-calibeating; specifically,*

$$0 \le \mathcal{B}^{\mathbf{c}}_t - \mathcal{R}^{\mathbf{b}}_t \le \gamma^2 |B| \frac{\ln t + 1}{t} \tag{11}$$

for all $t \ge 1$ and all sequences $\mathbf{a}_t \in A^t$ and $\mathbf{b}_t \in B^t$.

Proof. Our choice of $c_t = \bar{a}^{\mathbf{b}}_{t-1}(b_t)$ makes $\mathcal{B}^{\mathbf{c}}_t = \widetilde{\mathcal{R}}^{\mathbf{b}}_t$ for any $\mathbf{a}_t$ and $\mathbf{b}_t$; use Proposition 1 (see (7)).[33] □

The calibeating forecast c_t is thus the average of the actions in those periods $1 \le s \le t-1$ in which the forecast b_s was equal to the current forecast b_t. When B is a subset

[31] The restriction on the number of possible forecasts, i.e., on the number of bins, is needed in order for the resulting classification to be meaningful; in the extreme case where all forecasts are distinct, and thus each bin contains a single element, one gets $\mathcal{R}_t = 0$ for all t. This finiteness assumption may be relaxed; see Remark (d) below.

[32] If t is the first time that b_t is used (and so $\bar{a}^{\mathbf{b}}_{t-1}(b_t)$ is not defined), take c_t to be an arbitrary element of C.

[33] One always has $\widetilde{\mathcal{R}}^{\mathbf{b}}_t = \mathcal{B}^{\bar{\mathbf{a}}(\mathbf{b})}_t$; that is, the online refinement score $\widetilde{\mathcal{R}}^{\mathbf{b}}_t$ of the sequence $\mathbf{b} = (b_s)_{s \ge 1}$ is the same as the Brier score of the sequence of action averages $\bar{\mathbf{a}}(\mathbf{b}) = (\bar{a}_{s-1}(b_s))_{s \ge 1}$.

of C we get by (11) that[34] $\mathcal{B}_t^{\mathbf{c}} \le \mathcal{B}_t^{\mathbf{b}} - \mathcal{K}_t^{\mathbf{b}} + o(1)$; i.e., the Brier score of $\mathbf{c}$ is lower than that of $\mathbf{b}$ by essentially the calibration score of $\mathbf{b}$. Note that the specific values b_t of the B-forecasts are not used by the calibeating procedure ζ, and only the binning that they generate matters (see Section 2.3.1).

Remarks. *(a)* The simple calibeating procedure ζ is "universal" also in the sense of being independent of the specific set B: the forecast c_t is just the past average of the current bin.

(b) The history of one's own forecasts, $\mathbf{c}_{t-1}$, is not used by the procedure ζ; thus, c_t is a function of $\mathbf{a}_{t-1}$ and $\mathbf{b}_t$ only.

(c) One cannot guarantee a Brier score that is lower than the refinement score of $\mathbf{b}$. Indeed, for every $t \ge 1$ and every $\mathbf{b}_t \in B^t$, we have

$$\sup_{\mathbf{a}_t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \ge 0$$

for *any* sequence $\mathbf{c}_t$, because when all a_s are equal to a fixed $a^0 \in A$ we get $\mathcal{R}_t^{\mathbf{b}} = 0$ (because all bins contain only a^0, and so their variance is zero).

(d) If the set B is not finite, the procedure ζ calibeats also all sequences $\mathbf{b}$ with $N_t^{\mathbf{b}}/t \to 0$ as $t \to \infty$, where $N_t^{\mathbf{b}} := |\{b_s : s \le t\}|$ is the number of distinct forecasts used by $\mathbf{b}$ up to time t (use Proposition 1).

(e) From any forecasting procedure, whose forecasts may not be calibrated, we can generate by Theorem 3 another forecasting procedure that yields lower Brier scores in the long run, as follows. Let the $\mathbf{b}$-forecasts be generated by a forecasting procedure σ, and let σ' replace each b_t by the corresponding $\bar{a}_{t-1}^{\mathbf{b}}(b_t)$ (see Remark (b) in Appendix A.1.3 for some technical details); then σ' yields lower Brier scores than σ in the long run: $\mathcal{B}_t^{\mathbf{c}} \le \mathcal{B}_t^{\mathbf{b}} - \mathcal{K}_t^{\mathbf{b}} + o(1)$.

(f) The existence of a calibeating procedure may be proved by a minimax argument, extending the 1995 proof of Hart of calibration (see Section 4 of Foster and Vohra 1998, and the writeup of Hart 2021); we do so in Appendix A.2. The existence proof does not however provide an explicit calibeating procedure, for sure not the very simple one of Theorem 3.

Additional comments are relegated to Appendix A.1. In particular, we show that one cannot guarantee a calibeating error that is of an order of magnitude lower than $\log t/t$ (see Appendix A.1.1), and that the best that one can do is to decrease the error in Theorem 3 by a factor between 2 and 4 (depending on the dimension m), by using a more complex formula for the forecast c_t instead of (10) (see Appendix A.1.2).

[34] We use the standard notation: $f(x) = o(g(x))$ as $x \to x_0$ for $\lim_{x\to x_0} f(x)/g(x) = 0$; $f(x) = O(g(x))$ as $x \to x_0$ for $\limsup_{x\to x_0} f(x)/g(x) < \infty$; and $f(x) \sim g(x)$ as $x \to x_0$ for $\lim_{x\to x_0} f(x)/g(x) = 1$; in all of these x_0 could be infinity.

5 Self-Calibeating = Calibrating

The construction of Section 4 may be leveraged to obtain calibration. Indeed, when $\mathbf{b} = \mathbf{c}$ we have $\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}} = \mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{c}} = \mathcal{K}_t^{\mathbf{c}}$, and so "self-calibeating," i.e., $\mathbf{c}$ calibeating $\mathbf{c}$, is equivalent to calibration, i.e., $\mathcal{K}_t^{\mathbf{c}} \to 0$. To achieve this by the construct of Theorem 3 we would need to choose c_t so that $c_t = \bar{a}_{t-1}^{\mathbf{c}}(c_t)$. However, this requires a fixed point of the function $\bar{a}_{t-1}^{\mathbf{c}}(\cdot)$, which of course need not exist in general. We circumvent this by a "stochastic expected fixed point," i.e., by appealing to the corresponding "outgoing" theorems of Foster and Hart (2021)—see Appendix A.3 for details—and thereby obtain the classic calibration results (see Theorem 11(S) and (AD) in Foster and Hart 2021).[35]

Theorem 4 *Let $\delta > 0$ and let $D \subset C$ be a finite δ-grid of C. Then there exists a stochastic D-forecasting procedure σ of type MM that is δ-calibrated; specifically,*[36]

$$\mathbb{E}\left[\mathcal{K}_t\right] \leq \delta^2 + \gamma^2 |D| \frac{\ln t + 1}{t}$$

for all $t \geq 1$ and all sequences $\mathbf{a}_t \in A^t$. Moreover, σ may be taken to be δ-almost deterministic[37] *and then it is of type FP.*

Proof. For every t and history $h_{t-1} = (\mathbf{a}_{t-1}, \mathbf{c}_{t-1})$, the outgoing Theorem 11 (S) of Appendix A.3 applied to the function $\bar{a}_{t-1}(\cdot)$ yields a distribution η_t on D such that, using it as the distribution $\sigma(h_{t-1})$ of the forecast c_t, we have

$$\mathbb{E}_{t-1}\left[\|a_t - c_t\|^2 - \|a_t - \bar{a}_{t-1}(c_t)\|^2\right] \leq \delta^2 \tag{12}$$

for every $a_t \in A$, where $\mathbb{E}_{t-1}$ denotes expectation with respect to $\sigma(h_{t-1})$. Taking overall expectation and averaging over $t = 1, 2, ...$ yields $\mathbb{E}\left[\mathcal{B}_t - \widetilde{\mathcal{R}}_t\right] \leq \delta^2$; Proposition 1 completes the proof. For the moreover part, use part (AD) of Theorem 11. □

The proof is quite instructive: what we would like to get is $\lambda_t := \|a_t - c_t\|^2 - \|a_t - \bar{a}_{t-1}(c_t)\|^2 \leq 0$ no matter what a_t will be, which can be guaranteed only by choosing $c_t = \bar{a}_{t-1}(c_t)$. This means that c_t should be a fixed point of the function $\bar{a}_{t-1}(\cdot)$, a function that is far from being continuous,[38] and so need not in general have a fixed point. We thus use a distribution η_t instead—obtained by the minimax theorem—that guarantees that, in expectation, λ_t cannot exceed 0 by much (as in the simple illustration in Section 1.2 in Foster and Hart 2021).

[35] While the proof here may look different from the one in Foster and Hart (2021), the two proofs are in fact identical. The approach here with the online refinement score makes it more transparent.

[36] Since we are dealing here with only one forecasting sequence $\mathbf{c}$, we will drop the superscript $\mathbf{c}$ from $\mathcal{K}$ and $\bar{a}$.

[37] I.e., all randomizations are δ-local.

[38] It is defined only on the finite set $\{c_s : 1 \leq s \leq t\}$.

Remarks. *(a)* From inequality (12) for every history we get, by the Strong Law of Large Numbers for Dependent Random Variables (Loéve 1978, Theorem 32.1.E), that $\overline{\lim}_{t\to\infty}\left(\mathcal{B}_t - \widetilde{\mathcal{R}}_t\right) \leq \delta^2$ (a.s.), and thus $\overline{\lim}_{t\to\infty}\mathcal{K}_t \leq \delta^2$ (a.s.); see Appendix A5 in Foster and Hart (2021).

(b) Let D_t be an increasing sequence (i.e., $D_t \subseteq D_{t+1}$) of δ_t-grids of C such that $\delta_t \to 0$ and $|D_t|/t \to 0$ as $t \to \infty$; using D_t at time t guarantees that $\mathbb{E}[\mathcal{K}_t] = \mathbb{E}[\mathcal{B}_t - \mathcal{R}_t] = \mathbb{E}\left[\mathcal{B}_t - \widetilde{\mathcal{R}}_t\right] + \mathbb{E}\left[\widetilde{\mathcal{R}}_t - \mathcal{R}_t\right] \leq \delta_t^2 + O\left((|D_t|/t)\ln(t/|D_t|)\right) \to 0$ (by Proposition 1), and thus we obtain 0-calibration.

6 Calibeating by a Calibrated Forecast

While the procedure ζ of Section 4 calibeats any B-forecasting procedure, ζ itself need not yield calibrated forecasts (for example, if all its forecasts $c_t = \bar{a}_{t-1}^{\mathbf{b}}(b_t)$ are distinct, then all its bins are singletons and its calibration score is high), and so ζ itself may be calibeaten by yet another procedure. This suggests requiring our calibeating procedure to be calibrated, which is what we provide in this section.

Given two sequences $\mathbf{b}^1 = (b_t^1)_{t\geq 1}$ and $\mathbf{b}^2 = (b_t^2)_{t\geq 1}$ with values in sets B^1 and B^2, respectively, the resulting *joint binning* has $U = B^1 \times B^2$ as the set of bins; i.e., there is a (b^1, b^2)-bin for each pair $(b^1, b^2) \in B^1 \times B^2 = U$, and a_t is assigned to the u_t-bin where $u_t = (b_t^1, b_t^2)$. The bin averages are[39] $\bar{a}_t^{\mathbf{u}}(u) \equiv \bar{a}_t^{\mathbf{b}^1,\mathbf{b}^2}(b^1, b^2)$, and the refinement score is $\mathcal{R}_t^{\mathbf{u}} \equiv \mathcal{R}_t^{\mathbf{b}^1,\mathbf{b}^2} = (1/t)\sum_{s=1}^t \left\| a_s - \bar{a}_t^{\mathbf{b}^1,\mathbf{b}^2}(b_t^1, b_t^2)\right\|^2 \equiv (1/t)\sum_{s=1}^t \|a_s - \bar{a}_t^{\mathbf{u}}(v_t)\|^2$. Since $\mathcal{R}_t$ is the average internal variance of the bins, refining a binning—i.e., splitting bins into several new bins—can only decrease the refinement score;[40] see Appendix A.4 for a formal proof. Therefore

$$\mathcal{R}_t^{\mathbf{b}^1,\mathbf{b}^2} \leq \mathcal{R}_t^{\mathbf{b}^1} \quad \text{and} \quad \mathcal{R}_t^{\mathbf{b}^1,\mathbf{b}^2} \leq \mathcal{R}_t^{\mathbf{b}^2}. \tag{13}$$

By using the joint binning of the given sequence $\mathbf{b}$ together with our forecast $\mathbf{c}$, and appealing to the stochastic outgoing result, we obtain:

Theorem 5 *Let B be a finite set, and let $D \subset C$ be a finite δ-grid of C for some $\delta > 0$. Then there exists a stochastic $\mathbf{b}$-based D-forecasting procedure ζ of type MM that is (δ, B)-calibeating and δ-calibrated; specifically,*

$$\mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b},\mathbf{c}}\right] \leq \delta^2 + \gamma^2 |B|\,|D| \frac{\ln t + 1}{t},$$

[39] Formally,

$$\bar{a}_t^{\mathbf{u}}(x) := \frac{\sum_{1\leq s\leq t: u_s = x} a_s}{|\{1 \leq s \leq t : u_s = x\}|}$$

for every $x \in V$.

[40] To see why this is so, consider splitting a bin b with average $\bar{x}$ into two new bins b' and b'', with averages $\bar{x}'$ and $\bar{x}''$, respectively. Writing $\sum'$ and $\sum''$ for the sums over b' and b'', respectively, we have $\sum'(x_j - \bar{x}')^2 \leq \sum'(x_j - \bar{x})^2$ (this holds for any y in place of $\bar{x}$), and similarly for $\sum''$, which added together yields $\sum'(x_j - \bar{x}')^2 + \sum''(x_j - \bar{x}'')^2 \leq \sum(x_j - \bar{x})^2$.

and thus, by (13),

$$\begin{aligned}\mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] &\leq \delta^2 + \gamma^2|B|\,|D|\frac{\ln t + 1}{t} \quad \textit{and}\\ \mathbb{E}\left[\mathcal{K}_t^{\mathbf{c}}\right] &\leq \delta^2 + \gamma^2|B|\,|D|\frac{\ln t + 1}{t}\end{aligned}$$

for all $t \geq 1$ and all sequences $\mathbf{a}_t \in A^t$ and $\mathbf{b}_t \in B^t$. Moreover, ζ may be taken to be δ-almost deterministic and then it is of type FP.

Thus, ignoring the δ^2 term, in the long run the refinement score of ζ is no worse than that of any B-forecasting procedure, and its calibration score is zero. When $|B| = 1$ (and thus B-forecasting has no content), this reduces to the calibration result, Theorem 4, of Section 5.

Proof. At time t, given the history $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{b}_{t-1})$ together with b_t, apply Theorem 11 (S), respectively (AD), to the function $c \longmapsto \bar{a}_{t-1}^{\mathbf{b},\mathbf{c}}(b_t, c)$ (for $c \in D$) to get $\eta_t \in \Delta(D)$ such that, using it as the distribution of c_t given the history,[41] we have

$$\mathbb{E}_{t-1}\left[\|a_t - c_t\|^2 - \left\|a_t - \bar{a}_{t-1}^{\mathbf{b},\mathbf{c}}(c_t, b_t)\right\|^2\right] \leq \delta^2$$

for every $a_t \in A$. Taking overall expectation and averaging over t yields

$$\mathbb{E}\left[\mathcal{B}_t - \widetilde{\mathcal{R}}_t^{\mathbf{b},\mathbf{c}}\right] \leq \delta^2.$$

Proposition 1 completes the proof. □

Remarks. *(a)* Again, we can allow $N_t^{\mathbf{b}}$, the number of distinct forecasts used up to time t, to increase with t, provided that $N_t^{\mathbf{b}}/t \to 0$; see Remark (c) in Section 4 and Remark (b) in Section 5.

(b) The procedure in the proof above amounts to using a stochastic forecast-hedging calibration procedure separately in each bin in B.

(c) If the calibeating procedure of Theorem 3 is not calibrated, then one can construct another procedure that calibeats it, and then another one that calibeats that, and so on. A calibeating procedure that is calibrated, as obtained here, stops this infinite regress.[42]

(d) A minimax proof is provided in Appendix A.2.

[41] The resulting procedure is a **b**-based procedure, because the distribution η_t of c_t at time t is determined by the history $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{b}_{t-1})$ together with b_t. We write $\mathbb{E}_{t-1}$ for the resulting conditional expectation at time t (i.e., from the viewpoint of ζ).

[42] Which may well be quickly overwhelmed by the accumulating errors—of calibeating, as well as those due to getting a finite B by rounding up to a grid (see Remark (a) in Appendix A.1.3).

7 Calibeating by a Continuously Calibrated Forecast

In this section we show that one can guarantee calibeating by a *deterministic* procedure that is *continuously calibrated*, a useful weakening of calibration defined in Foster and Hart (2021).

Let us recall the definition of continuous calibration. A *(fractional) binning* is a collection $\Pi = (w_i)_{i \in I}$ of weight functions $w_i : C \to [0,1]$ for $i \in I$ such that $\sum_{i \in I} w_i(c) = 1$ for all $c \in C$, where I is a finite or countably infinite set; the binning Π is *continuous* if all the w_i are continuous functions on C. The interpretation is that at each period s the fraction $w_i(c_s)$ of $z_s = a_s - c_s$ is assigned to each bin i in I. A deterministic forecasting procedure σ is *continuously calibrated* if

$$\lim_{t \to \infty} \left(\sup_{\mathbf{a}_t} \mathcal{K}_t^{\Pi} \right) = 0 \tag{14}$$

for every continuous binning Π, where the Π-*calibration* score $\mathcal{K}_t^{\Pi}$ is[43]

$$\mathcal{K}_t^{\Pi} := \sum_{i \in I} \left(\frac{n_t^i}{t} \right) \left\| e_t^i \right\|^2$$

with $n_t^i := \sum_{s=1}^t w_i(c_s)$ and $e_t^i := \sum_{s=1}^t (w_i(c_s)/n_t^i)(a_s - c_s)$. Proposition 3 in Foster and Hart (2021) shows that it suffices to require (14) for one specific continuous binning Π_0; i.e., σ is continuously calibrated if and only if (14) holds for $\Pi = \Pi_0$.

Let B be an arbitrary finite set[44] and let $\Pi = (w_i)_{i \in I}$ be a fractional binning. Consider the joint fractional binning with bins $U := B \times I$, where at each time t the fractions $w_i(c_t)$ of $a_t - c_t$ are assigned to bins (b_t, i) for all $i \in I$; that is, each bin $(b, i) \in B \times I$ gets the fraction

$$\lambda_t(b, i) := \mathbf{1}_b(b_t) w_i(c_t),$$

where $\mathbf{1}_x$ stands for the x-indicator function (i.e., $\mathbf{1}_x(y) = 1$ for $y = x$ and $\mathbf{1}_x(y) = 0$ for $y \neq x$). Consider bin (b, i) at time t; its total weight, average, and variance are,

[43] A more precise, but cumbersome, notation would be $\mathcal{K}^{\Pi(\mathbf{c})}$, since at each time t the binning is given by $\Pi(c_t) = (w_i(c_t))_{i \in I}$.

[44] One may easily generalize to fractional B binnings; also, B could be infinite when the binning is continuous (or, more generally, when the binning is uniformly approximable by finite fractional binnings, as in (9) in Foster and Hart 2021).

respectively,

$$\begin{aligned} n_t(b,i) &:= \sum_{s=1}^{t} \lambda_s(b,i), \\ e_t(b,i) &:= \sum_{s=1}^{t} \left(\frac{\lambda_s(b,i)}{n_t(b,i)}\right)(a_s - c_s), \quad \text{and} \\ v_t(b,i) &:= \sum_{s=1}^{t} \left(\frac{\lambda_s(b,i)}{n_t(b,i)}\right)\|a_s - c_s - e_t(b,i)\|^2 \end{aligned}$$

(note that $n_t(b,i) = \sum_{s\le t: b_s = b} w_i(c_s)$). The calibration and refinement scores are then

$$\begin{aligned} \mathcal{K}_t^{\mathbf{b},\Pi} &:= \sum_{(b,i)\in B\times I} \left(\frac{n_t(b,i)}{t}\right)\|e_t(b,i)\|^2 \quad \text{and} \\ \mathcal{R}_t^{\mathbf{b},\Pi} &:= \sum_{(b,i)\in B\times I} \sum_{s=1}^{t} \left(\frac{\lambda_s(b,i)}{t}\right) v_t(b,i). \end{aligned}$$

We can now state the result for continuous calibration.

Theorem 6 *Let B be a finite set. Then there exists a deterministic $\mathbf{b}$-based forecasting procedure ζ of type FP that is B-calibeating and is continuously calibrated; specifically,*

$$\mathcal{B}_t^{\mathbf{c}} \le \mathcal{R}_t^{\mathbf{b},\Pi_0} + o(1), \tag{15}$$

where Π_0 is the continuous binning given by Proposition 3 in Foster and Hart (2021); this implies that

$$\mathcal{B}_t^{\mathbf{c}} \le \mathcal{R}_t^{\mathbf{b}} + o(1),$$

and that ζ is continuously calibrated. All these hold as $t \to \infty$ uniformly over all sequences $\mathbf{a}$ and $\mathbf{b}$.

To prove this we use the corresponding *online refinement* score $\widetilde{\mathcal{R}}_t^{\mathbf{b},\Pi}$, in which the offline average e_t is replaced with the online average e_{s-1}; namely,

$$\widetilde{\mathcal{R}}_t^{\mathbf{b},\Pi} := \sum_{(b,i)\in B\times I} \sum_{s=1}^{t} \left(\frac{\lambda_s(b,i)}{t}\right)\|a_s - c_s - e_{s-1}(b,i)\|^2 .$$

The parallel result to Proposition 1 is

Proposition 7 *For every finite set B and every continuous binning $\Pi = (w_i)_{i=1}^I$ on C, as $t \to \infty$ we have*

$$0 \le \widetilde{\mathcal{R}}_t^{\mathbf{b},\Pi} - \mathcal{R}_t^{\mathbf{b},\Pi} \le o(1)$$

uniformly over all sequences $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$.

The proof is an adaptation of the proof of Proposition 1 to fractional binnings, and is relegated to Appendix A.6.

Proof of Theorem 6. Take Π to be $\Pi_0 = (w_i)_{i \in I}$ of Proposition 3 in Foster and Hart (2021). At time t, given $\mathbf{a}_{t-1}, \mathbf{c}_{t-1}$, and $\mathbf{b}_t$, the Outgoing Fixed Point Theorem, specifically, Theorem 11 (D), applied to the continuous function $c \longmapsto c - \sum_{i \in I} w_i(c) e_{t-1}(b_t, i)$, yields $c_t \in C$ such that

$$\begin{aligned} \|a_t - c_t\|^2 &\leq \left\| a_t - c_t - \sum_{i \in I} w_i(c_t) e_{t-1}(b_t, i) \right\|^2 \\ &\leq \sum_{i \in I} w_i(c_t) \left\| a_t - c_t - e_{t-1}(b_t, i) \right\|^2 \end{aligned}$$

for every $a_t \in A$ (the second inequality is by the convexity of $\|\cdot\|^2$). Averaging over t gives $\mathcal{B}_t \leq \widetilde{\mathcal{R}}_t^{\mathbf{b}, \Pi_0}$, and thus (15) by Proposition 7. To complete the proof use $\mathcal{R}_t^{\mathbf{b}, \Pi_0} \leq \mathcal{R}_t^{\mathbf{b}}$ and $\mathcal{R}_t^{\mathbf{b}, \Pi_0} \leq \mathcal{R}_t^{\Pi_0}$ by the refinement monotonicity of the refinement score (see Appendix A.4), and then $\mathcal{K}_t^{\Pi_0} = \mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\Pi_0}$ by the generalization of the Brier score decomposition to fractional binnings (see Appendix A.5). □

8 Multi-Calibeating

Suppose that there are $N \geq 1$ forecasting sequences, $\mathbf{b}^n = (b_t^n)_{t \geq 1}$ for $n = 1, 2, ..., N$. We assume that each $\mathbf{b}^n$ uses only finitely many forecasts: there is a finite set B^n such that $b_t^n \in B^n$ for all $t \geq 1$ (and, as in Section 2.3.1, while B^n could be a subset of C, it may well be an arbitrary set). Put $\mathbf{b} = (\mathbf{b}^1, ..., \mathbf{b}^N)$; we are looking for a $\mathbf{b}$-based forecasting procedure—i.e., c_t is determined after all the $b_t^1, ..., b_t^N$ are announced (and hence is a function of $\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{b}_t^1, ..., \mathbf{b}_t^N$—that simultaneously calibeats all the $\mathbf{b}^n$ sequences. We have:

Theorem 8 *(i) There exists a simple deterministic $(\mathbf{b}^1, ..., \mathbf{b}^N)$-based forecasting procedure ζ that is B^n-calibeating for all $n = 1, ..., N$; specifically, the forecast of ζ in period t is $c_t = \bar{a}_{t-1}^{\mathbf{b}^1, ..., \mathbf{b}^N}(b_t^1, ..., b_t^N)$, the average of the actions in all past periods $s \leq t-1$ where the combination $(b_t^1, ..., b_t^N)$, was used.*[45]

(ii) For every finite δ-grid D of C, there exists a stochastic $(\mathbf{b}^1, ..., \mathbf{b}^N)$-based D-forecasting procedure ζ of type MM that is (δ, B^n)-calibeating for all $n = 1, ..., N$, and is δ-calibrated.

(iii) For every finite δ-grid D of C, there exists a stochastic $(\mathbf{b}^1, ..., \mathbf{b}^N)$-based D-forecasting procedure ζ of type FP that is (δ, B^n)-calibeating for all $n = 1, ..., N$, is δ-calibrated, and is δ-almost deterministic.

[45] Again, if t is the first period in which $(b_t^1, ..., b_t^N)$ is used, take $c_t \in C$ to be arbitrary.

(iv) There exists a deterministic $(\mathbf{b}^1, ..., \mathbf{b}^N)$*-based* C*-forecasting procedure* ζ *of type FP that is* (δ, B^n)*-calibeating for all* $n = 1, ..., N$*, and is continuously calibrated.*

Proof. This is immediate from the results of the previous sections by taking $(\mathbf{b}^1, ..., \mathbf{b}^N)$ as $\mathbf{b}$ and using inequalities such as $\mathcal{R}_t^{\mathbf{b}^1,...,\mathbf{b}^N} \leq \mathcal{R}_t^{\mathbf{b}^n}$ for each n by Appendix A.4. □

Remarks. *(a)* The error term is

$$\gamma^2 \prod_{n=1}^{N} |B^n| \frac{\ln t + 1}{t};$$

thus, in (i) we have

$$\mathcal{B}_t^{\mathbf{c}} \leq \mathcal{R}_t^{\mathbf{b}^n} + \gamma^2 \prod_{n=1}^{N} |B^n| \frac{\ln t + 1}{t}, \tag{16}$$

and in (ii) we have

$$\begin{aligned} \mathbb{E}\left[\mathcal{K}_t^{\mathbf{c}}\right] &\leq \delta^2 + \gamma^2 |D| \prod_{n=1}^{N} |B^n| \frac{\ln t + 1}{t} \quad \text{and} \\ \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}}\right] &\leq \mathbb{E}\left[\mathcal{R}_t^{\mathbf{b}^j}\right] + \delta^2 + \gamma^2 |D| \prod_{n=1}^{N} |B^n| \frac{\ln t + 1}{t} \end{aligned}$$

for all $n = 1, ..., N$, all $t \geq 1$, and all sequences $\mathbf{a}, \mathbf{b}^1, ..., \mathbf{b}^N$.

(b) Since the constant $\prod_{n=1}^{N} |B^n|$ in the above error terms increases exponentially with N, we provide in Appendix A.7 multi-calibeating procedures that are more complex but yield smaller error terms.

(c) One may again allow the B^n to be infinite, provided that $N_t^{\mathbf{b}^n}/t \to 0$ as $t \to \infty$.

(d) The result in (iii) holds with probability one, and not only in expectation; see Remark (a) in Section 6 and Appendix A5 in Foster and Hart (2021).

A Appendix

The appendix contains additional remarks and results.

A.1 A Simple Way to Calibeat

We consider here the minimal calibeating error that can be guaranteed. First, we show in Appendix A.1.1 that it must be at least of the order of $\log t/t$, the same order obtained by Theorem 3; second, we pin down the constant in Appendix A.1.2: it is within a factor between 2 and 4 (depending on the dimension m and the geometric shape of the set C) of the constant of Theorem 3. Additional comments on Section 4 are provided in A.1.3.

A.1.1 The Calibeating Error

We prove here that one cannot guarantee a calibeating error of an order of magnitude lower than $\log t/t$. We show that this is so already in the simplest one-dimensional case; see Remark (a) below for the extension to the multidimensional case.

Proposition 9 *Let $A = \{0,1\}$ and $C = [0,1]$, and let $\mathbf{b}$ be a constant sequence (e.g., $b_t = 1/2$ for all t). Then for every $\mathbf{b}$-based forecasting procedure ζ we have*[46]

$$\sup_{\mathbf{a}_t \in A^t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \geq \left(\frac{1}{4} - o(1)\right) \frac{\ln t}{t} \tag{17}$$

as $t \to \infty$.

Proof. Consider the game between the "action player" who chooses the actions a_t and the "calibrating player" who chooses the sequence of forecasts c_t, with payoff $\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}$ (cf. the "calibration game" in Foster and Hart 2018, where the payoff was $\mathcal{K}_t^{\mathbf{c}}$). We will provide a mixed strategy of the action player that guarantees that

$$\inf_{\zeta} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \geq \left(\frac{1}{4} - o(1)\right) \frac{\ln t}{t}, \tag{18}$$

where the infimum is taken over all forecasting procedures[47] ζ. This implies that for every such ζ there is for each $t \geq 1$ at least one sequence $\mathbf{a}_t$ in A^t for which the same inequality holds; this is (17).

The mixed strategy of the action player that we provide consists of conditionally i.i.d. actions, specifically, $a_t|\theta \sim \text{Bernoulli}(\theta)$ where $\theta \sim \text{Beta}(\alpha, \alpha)$ for a fixed $\alpha > 0$ (this is the so-called "beta-binomial" distribution with parameters $\alpha = \beta$). The following formulas are well known, and easy to see (e.g., Johnson, Kemp, and Kotz 2005):

$$\mathbb{E}\left[\bar{a}_t\right] = \frac{1}{2} \quad \text{and} \quad \mathbb{V}ar\left[\bar{a}_t\right] = \mathbb{V}ar\left[\bar{a}_t\right] = \frac{t + 2\alpha}{4(2\alpha+1)t}; \tag{19}$$

the Bayesian estimate of θ given the history h_t is

$$\hat{\theta}_t := \mathbb{E}\left[\theta|h_t\right] = \frac{t\bar{a}_t + \alpha}{t + 2\alpha};$$

thus, by (19),

$$\mathbb{E}\left[\hat{\theta}_t\right] = \frac{1}{2} \quad \text{and} \quad \mathbb{V}ar\left[\hat{\theta}_t\right] = \frac{t}{4(2\alpha+1)(t+2\alpha)}. \tag{20}$$

[46] For a constant sequence $\mathbf{b}$, a $\mathbf{b}$-based forecasting procedure is simply a forecasting procedure. The expectation in (17) is over the stochastic choices of ζ (and it applies only to $\mathcal{B}_t^{\mathbf{c}}$, since $\mathcal{R}_t^{\mathbf{b}}$ is determined by $\mathbf{a}_t$ alone when $\mathbf{b}_t$ is a constant sequence).

[47] The expectation in (18) is over the randomizations of both actions and forecasts.

The sequence $\mathbf{b}$ yields a single bin, and so $\mathcal{R}_t^{\mathbf{b}}$ is the variance of $a_1, ..., a_t$, i.e., $\bar{a}_t(1-\bar{a}_t)$, which, by (19), gives

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_t^{\mathbf{b}}\right] &= \mathbb{E}\left[\bar{a}_t(1-\bar{a}_t)\right] = \mathbb{E}\left[\bar{a}_t\right] - \mathbb{E}^2\left[\bar{a}_t\right] - \mathbb{V}ar\left[\bar{a}_t\right] \\ &= \frac{1}{2} - \frac{1}{4} - \frac{t+2\alpha}{4(2\alpha+1)t} = \lambda - \frac{\lambda}{t}, \end{aligned} \tag{21}$$

where we used $\mathbb{E}\left[\bar{a}_t^2\right] = \mathbb{E}^2\left[\bar{a}_t\right] + \mathbb{V}ar\left[\bar{a}_t\right]$, and put

$$\lambda := \frac{\alpha}{2(2\alpha+1)}.$$

Next, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}}\right] &= \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[(a_s - c_s)^2\right] = \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{E}\left[(a_s - c_s)^2 | h_{s-1}\right]\right] \\ &\geq \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{V}ar\left[a_s | h_{s-1}\right]\right] = \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\left[\hat{\theta}_{s-1}(1-\hat{\theta}_{s-1})\right], \end{aligned}$$

where the inequality is by $\mathbb{E}\left[(X-Y)^2\right] \geq \mathbb{V}ar[X]$ for any Y that is independent of[48] X, and the equality following it is by $a_s|h_{s-1} \sim \text{Bernoulli}(\theta|h_{s-1}) = \text{Bernoulli}(\hat{\theta}_{s-1})$. Using (20) we get

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_{s-1}(1-\hat{\theta}_{s-1})\right] &= \mathbb{E}\left[\hat{\theta}_{s-1}\right] - \mathbb{E}^2\left[\hat{\theta}_{s-1}\right] - \mathbb{V}ar\left[\hat{\theta}_{s-1}\right] \\ &= \frac{1}{2} - \frac{1}{4} - \frac{s-1}{4(2\alpha+1)(s-1+2\alpha)} = \lambda + \frac{\lambda}{s+2\alpha-1}, \end{aligned}$$

and thus

$$\mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}}\right] \geq \lambda + \frac{\lambda}{t}\sum_{s=1}^{t}\frac{1}{s+2\alpha-1}.$$

Together with (21) this yields

$$\mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \geq \frac{\lambda}{t}\sum_{s=1}^{t}\frac{1}{s+2\alpha-1} + \frac{\lambda}{t} \sim \lambda\frac{\ln t}{t}$$

as $t \to \infty$. Since λ can be made arbitrarily close to[49] $1/4$ by taking large enough α we get (18), which completes the proof. □

[48] Use $\mathbb{E}\left[(X-y)^2\right] \geq \mathbb{V}ar\left[X\right]$ for each value y of Y. The inequality holds more generally for nonpositively correlated X and Y, since $\mathbb{E}\left[(X-Y)^2\right] \geq \mathbb{V}ar\left[X-Y\right] = \mathbb{V}ar\left[X\right] - 2\,\mathbb{C}ov\left[X,Y\right] + \mathbb{V}ar\left[Y\right]$, which is $\geq \mathbb{V}ar\left[X\right]$ when $\mathbb{C}ov\left[X,Y\right] \leq 0$.

[49] As $\alpha \to \infty$ the beta-binomial distribution converges to the binomial distribution with $\theta = 1/2$, for which $\mathcal{R}_t \approx 1/4$. We cannot however use this limit distribution, since θ being fixed yields a much smaller error, of the order of $1/t$ instead of $\log t/t$.

Remarks. *(a)* In the multidimensional case with $A = \{0,1\}^m$ and $C = [0,1]^m$ (for any $m \geq 1$), let $\mathbf{b}$ be a constant sequence; applying the above result to each one of the m coordinates separately and then summing up yields

$$\sup_{\mathbf{a}_t \in A^t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \geq \left(\frac{m}{4} - o(1)\right) \frac{\ln t}{t} \tag{22}$$

as $t \to \infty$.

(b) Given a finite set B, let the sequence $\mathbf{b}$ use all b in B with equal frequencies (for example, let the b_t alternate in a round-robin manner between the elements of B); applying Proposition 9 to the subsequence where $b_t = b$ for each $b \in B$ separately and then summing up yields

$$\sup_{\mathbf{a}_t \in A^t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \geq \left(\frac{1}{4} - o(1)\right) |B| \, \frac{\ln(t/|B|)}{t}.$$

A.1.2 Improving the Constant

Here we will show that one can improve the calibeating error of Theorem 3 by a factor between 2 and 4 (depending on the dimension and the shape of the set C), and this essentially matches the lower bound of the previous Section A.1.1.

Assuming that C is a full-dimensional set in[50] $\mathbb{R}^m$, let r be the *radius of the minimal bounding sphere of* C; thus, r is minimal such that $C \subseteq \bar{B}(c^0, r)$ for some $c^0 \in C$. The relation of r to the diameter γ of C is, by Jung's (1901) theorem,

$$r^2 \leq \gamma^2 \frac{m}{2(m+1)} \tag{23}$$

(and, of course, $\gamma \leq 2r$).

Proposition 10 *Let B be a finite set, and let ζ' be the deterministic $\mathbf{b}$-based forecasting procedure given by*

$$c'_t := \left(1 - \frac{1}{n_t^{\mathbf{b}}(b_t)}\right) \bar{a}_{t-1}^{\mathbf{b}}(b_t) + \frac{1}{n_t^{\mathbf{b}}(b_t)} c^0$$

for every time $t \geq 1$. Then ζ is B-calibeating, and

$$\mathcal{B}_t^{\mathbf{c}'} - \mathcal{R}_t^{\mathbf{b}} \leq r^2 |B| \frac{\ln t + 1}{t} \tag{24}$$

for all $t \geq 1$ and all sequences $\mathbf{a}_t \in A^t$ and $\mathbf{b}_t \in B^t$.

Thus, the forecast c'_t of ζ' is an appropriately weighted average of the forecast $c_t = \bar{a}_{t-1}^{\mathbf{b}}(b_t)$ of the procedure ζ of Theorem 3 and the fixed "center" point c^0 of C. Compared with (11), the upper bound of (24) on the calibeating error has r^2 instead of γ^2, which,

[50] If the affine space spanned by $C \subset \mathbb{R}^m$ has a lower dimension $m' < m$, project everything to $\mathbb{R}^{m'}$.

by (23), is an improvement by a factor of at least 2; when $m = 1$, by a factor of 4. Of course, ζ' gives up somewhat on the extreme simplicity of ζ, i.e., (10).

Proof. For any vectors $x, y \in \mathbb{R}^m$ and any scalar $\nu \in [0,1]$, we have $\|x - (1-\nu)y\|^2 - (1-\nu)\|x-y\|^2 = \nu\|x\|^2 - \nu(1-\nu)\|y\|^2 \leq \nu\|x\|^2$. Applying this to $x = a_s - c^0$, $y = \bar{a}^{\mathbf{b}}_{s-1}(b_s) - c^0$, and $\nu = 1/n^{\mathbf{b}}_s(b_s)$ yields

$$\|a_s - c'_s\|^2 - \left(1 - \frac{1}{n^{\mathbf{b}}_s(b_s)}\right)\left\|a_s - \bar{a}^{\mathbf{b}}_{s-1}(b_s)\right\|^2 \leq \frac{1}{n^{\mathbf{b}}_s(b_s)}\left\|a_s - c^0\right\|^2 \leq \frac{r^2}{n^{\mathbf{b}}_s(b_s)}.$$

Averaging the left-hand side for $s = 1, ..., t$ yields $\mathcal{B}^{\mathbf{c}'}_t - \mathcal{R}^{\mathbf{b}}_t$, and so, putting $B_t := \{b \in B : n^{\mathbf{b}}_t(b) > 0\}$, we get

$$\mathcal{B}^{\mathbf{c}'}_t - \mathcal{R}^{\mathbf{b}}_t \leq \frac{1}{t}\sum_{b \in B_t}\sum_{i=1}^{n^{\mathbf{b}}_t(b)}\frac{r^2}{i} \leq \frac{1}{t}|B|\, r^2(\ln t + 1)$$

(because $|B_t| \leq |B|$ and $n^{\mathbf{b}}_t(b) \leq t$), which is (24). □

Remark. When $C = [0,1]^m$ for some $m \geq 1$, we have $r^2 = m/4$ (take $c^0 = (1/2, ..., 1/2)$), and thus

$$\mathcal{B}^{\mathbf{c}'}_t - \mathcal{R}^{\mathbf{b}}_t \leq \frac{m|B|(\ln t + 1)}{4t}.$$

When B is a singleton and $\mathbf{b}$ is a constant sequence, this upper bound is $(m/4 + o(1))\ln t/t$, which is asymptotically the same as the lower bound of (22).

A.1.3 Additional Comments

We provide here a number of additional remarks on the result of Theorem 3 on simple calibeating.

Remarks. *(a)* If we want our forecasts c_t to lie in a given δ-grid $D \subset C$ (which may or may not be the same as the grid B used for $\mathbf{b}$), then taking $c_t \in D$ to be within δ of $\bar{a}^{\mathbf{b}}_{t-1}(b_t)$ introduces an additional error of $2\gamma\delta$ in (11) (because $\left|\|a-c\|^2 - \|a-d\|^2\right| \leq 2\gamma\|d-c\|$ for all $a, c, d \in C$), and thus it yields $(\sqrt{2\gamma\delta}, B)$-calibeating.

(b) Let the $\mathbf{b}$-forecasts be generated by a forecasting procedure σ; the procedure σ' of Remark (e) in Section 4, whereby each b_t is replaced by the corresponding $\bar{a}^{\mathbf{b}}_{t-1}(b_t)$, and which guarantees a lower Brier score than σ in the long run, is implemented as follows. In each period t one computes the forecast b_t according to σ, and then announces $c_t = \bar{a}^{\mathbf{b}}_{t-1}(b_t)$ (the b_t is not announced). To carry this out one needs to recall the history $\mathbf{b}_{t-1}$, which in general need not be deducible from the history $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1})$ of σ' (because different b-bins may have had the same average, and so different b_s may have yielded the same $c_s = \bar{a}^{\mathbf{b}}_{s-1}(b_s)$). In game-theoretic terms, the resulting σ' is *not* a *behavior* strategy

(which is what we have defined a forecasting procedure to be, in Section 2.2), but rather a *mixed* strategy (i.e., a probabilistic mixture of pure, deterministic, strategies). However, since the game between the "forecasting player" and the "action player" is a game of perfect recall, by Kuhn's (1953) theorem the mixed strategy σ' induces an equivalent behavior strategy σ'', which is thus a forecasting procedure (this "equivalence" means that no matter what the action player does, the probability of any outcome is the same under the mixed strategy and the induced behavior strategy). The construction of σ'' is straightforward (see, e.g., Hart 1992): for every $t \geq 1$, history $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}) \in A^{t-1} \times C^{t-1}$, and forecast $c_t \in C$, let $\Gamma_{t-1} := \{\mathbf{b}_{t-1} : \bar{a}^{\mathbf{b}}_{s-1}(b_s) = c_s \text{ for every } 1 \leq s \leq t-1\}$ and $\Gamma_t := \{\mathbf{b}_t : \bar{a}^{\mathbf{b}}_{s-1}(b_s) = c_s \text{ for every } 1 \leq s \leq t\}$ be the sets of $\mathbf{b}_{t-1}$ and $\mathbf{b}_t$ histories that, together with the given $\mathbf{a}_{t-1}$, yield $\mathbf{c}_{t-1}$ and $\mathbf{c}_t$, respectively; then the probability that σ'' forecasts c_t after $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1})$ is given by

$$\sigma''(\mathbf{a}_{t-1}, \mathbf{c}_{t-1})(c_t) := \frac{\sum_{\mathbf{b}_t \in \Gamma_t} \prod_{s=1}^{t} \sigma(\mathbf{a}_{s-1}, \mathbf{b}_{s-1})(b_s)}{\sum_{\mathbf{b}_{t-1} \in \Gamma_{t-1}} \prod_{s=1}^{t-1} \sigma(\mathbf{a}_{s-1}, \mathbf{b}_{s-1})(b_s)}. \tag{25}$$

(c) Can one do better than by choosing $c_t = \bar{a}^{\mathbf{b}}_{t-1}(b_t)$ at each time t? Since $\widetilde{\mathcal{R}}^{\mathbf{b}}_t - \mathcal{R}^{\mathbf{b}}_t = O(\log t/t) \to 0$, consider the game where our $\mathbf{c}$-forecaster wants to minimize $\mathcal{B}^{\mathbf{c}}_t - \widetilde{\mathcal{R}}^{\mathbf{b}}_t$ (instead of $\mathcal{B}^{\mathbf{c}}_t - \mathcal{R}^{\mathbf{b}}_t$) against an opponent that controls the sequences $\mathbf{a}_t$ and $\mathbf{b}_t$; alternatively, the opponent controls the sequence $\mathbf{a}_t$, whereas the sequence $\mathbf{b}_t$ is exogenous or is determined by history (i.e., by a forecasting procedure). We claim that the strategy ζ of Theorem 3 is the *unique* subgame-perfect optimal strategy of our forecaster. To see this, consider

$$\mathcal{B}^{\mathbf{c}}_t - \widetilde{\mathcal{R}}^{\mathbf{b}}_t = \frac{1}{t} \sum_{s=1}^{t} \left[||a_s - c_s||^2 - \left\| a_s - \bar{a}^{\mathbf{b}}_{s-1}(b_s) \right\|^2 \right]. \tag{26}$$

Suppose that we are at a period $r \leq t$, and so the terms $s < r$ of the sum (26) are all given. To *guarantee*—no mater what the future a_s and b_s will be—that the sum of the remaining terms, i.e., $s \geq r$, is as small as possible one *must* choose now $c_r = \bar{a}^{\mathbf{b}}_{r-1}(b_r)$. This follows since for every $\bar{a} \in \operatorname{conv} A$ and $c \neq \bar{a}$ we have[51]

$$\sup_{a \in A} [\|a - c\|^2 - \|a - \bar{a}\|^2] \geq \|c - \bar{a}\|^2 > 0,$$

[51] One way to see this is as follows. Let $\bar{a} = \sum_i \lambda_i a_i$ for $a_i \in A$, $\lambda_i > 0$, $\sum_i \lambda_i = 1$; then

$$\sum_i \lambda_i \|a_i - c\|^2 = \sum_i \lambda_i \|a_i - \bar{a}\|^2 + \|c - \bar{a}\|^2$$

(because $\bar{a}$ is the weighted average of the a_i), and so for some a_i we must have

$$\|a_i - c\|^2 \geq \|a_i - \bar{a}\|^2 + \|c - \bar{a}\|^2.$$

whereas $\|a-c\|^2 - \|a-\bar{a}\|^2 = 0$ for every a when $c = \bar{a}$. Thus

$$\min_{c\in C} \sup_{a\in A} [\|a-c\|^2 - \|a-\bar{a}\|^2] = 0,$$

with a *unique* minimizer at $c = \bar{a}$. Therefore, for any sequence $\mathbf{b}_t$ we have

$$\min_{c_r,...,c_t\in C} \sup_{a_r,...,a_t\in A} \sum_{s=r}^{t} \left[||a_s - c_s||^2 - \left\|a_s - \bar{a}_{s-1}^{\mathbf{b}}(b_s)\right\|^2 \right] = 0,$$

with the minimum *uniquely* attained by choosing $c_s = \bar{a}_{s-1}^{\mathbf{b}}(b_s)$ for each $s = r, ..., t$.

A.2 A Minimax Proof of Calibeating

The simplest proof of the existence of forecasts that are guaranteed to be calibrated consists of an application of the von Neumann's (1928) minimax theorem for finite two-person zero-sum games; see Hart (2021).[52] We provide here a minimax proof of calibeating as well. Note that these minimax proofs do not yield actual constructions of the corresponding procedures; they just give simple proofs of their existence.

The *calibeating game* consists of two players, which we call the "AB-player" and the "C-player". In each period $s = 1, 2, ..., t$ the AB-player chooses a pair $(a_s, b_s) \in A \times B$, then the chosen b_s is revealed, and finally the C-player chooses a forecast[53] $c_s \in C$. At period t, when the game ends, the payoff of the C-player is his calibration score $\mathcal{K}_t^{\mathbf{c}}$. The joint choice of the action a_s and the forecast b_s is equivalent to allowing an arbitrary dependence between them; i.e., the **b**-forecaster may have any degree of "knowledge" or "expertise" about the action sequence **a** (from "no knowledge," where a_s and b_s are chosen independently, all the way to "complete knowledge," where they are fully correlated, e.g., $b_s = a_s$).

The sets A and B are assumed to be finite (with $A \subset C$), and we will restrict the C-player to use a finite set $D \subset C$, which makes the calibeating game a finite game. Moreover, we will make the **c**-binning a refinement of the **b**-binning, and so $\mathcal{R}_t^{\mathbf{c}} \leq \mathcal{R}_t^{\mathbf{b}}$ (by Proposition 12 in Appendix A.4 below). Therefore, a strategy σ of the C-player that guarantees $\mathbb{E}[\mathcal{K}_t^{\mathbf{c}}] \leq \varepsilon$ against any strategy τ of the AB-player gives $\mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] = \mathbb{E}\left[\mathcal{K}_t^{\mathbf{c}} + \mathcal{R}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \leq \mathbb{E}[\mathcal{K}_t^{\mathbf{c}}] \leq \varepsilon$, and thus it yields B-calibeating, moreover by a calibrated procedure (cf. Theorem 5).

We proceed as follows. For every $b \in B$ we take D_b to be a finite δ-grid of C, such that these grids are disjoint, i.e., $D_b \cap D_{b'} = \emptyset$ for all $b \neq b'$; put $D := \cup_{b\in B} D_b$. We then

[52] Which is a writeup of a proof provided in 1995; see Section 4, "An Argument of Sergiu Hart," in Foster and Vohra (1998). For generalizations see Sandroni (2003), Olszewski and Sandroni (2008), Shmaya (2008).

[53] Equivalently: first b_s is chosen by the AB-player and is publicly announced, and then a_s and c_s are chosen by the two players independently.

restrict the C-player to use only forecasts in D_b after b is announced, i.e., $c_s \in D_{b_s}$ for every s; this indeed makes the $\mathbf{c}$-binning a refinement of the $\mathbf{b}$-binning.

We claim that for every mixed strategy τ of the AB-player there is a strategy σ of the C-player such that $\mathbb{E}[\mathcal{K}_t^{\mathbf{c}}] \leq \varepsilon := O(\delta + 1/\sqrt{t})$ (we do not try optimize the bounds here). Indeed, in each period $s \leq t$ choose $c_s \in D_{b_s}$ to be such that $\|c_s - \mathbb{E}[a_s | h_{s-1}, b_s]\| \leq \delta$ (i.e., take the conditional probability—given the history $h_{s-1} = (\mathbf{a}_{s-1}, \mathbf{b}_{s-1}, \mathbf{c}_{s-1})$ and the announced b_s—that is generated by the mixed strategy τ of the AB-player, and round it up to the δ-grid D_{b_s} that is used for b_s). A standard computation, as in Hart (2021), shows that $\mathbb{E}[K_t] \leq \delta + \gamma\sqrt{|D|}/\sqrt{t}$, where $K_t := (1/t)\sum_{d \in D} n_t(d)\,\|e_t(d)\|$ is the ℓ_1-calibration score, for which we trivially have $\mathcal{K}_t \leq \gamma K_t$ (see footnote 15); this proves the claim.

The minimax theorem therefore yields a strategy σ of the C-player that guarantees $\mathbb{E}[\mathcal{K}_t^{\mathbf{c}}] \leq \varepsilon$ against all strategies τ of the AB-player; we have thus obtained calibration and, as shown above, B-calibeating.

A.3 "Outgoing" Results

We provide here the results of the "outgoing" theorems of Foster and Hart (2021), stating them in a convenient manner for our purpose.[54] A probability distribution η is called "δ-local" if its support is included in a ball of radius δ; i.e., there is y such that $\eta(\bar{B}(y;\delta)) = 1$.

Theorem 11 *Let $C \subset \mathbb{R}^m$ be a nonempty compact convex set.*

(D) *Let $g : C \to \mathbb{R}^m$ be a continuous function. Then there exists a point y in C that is of type FP, such that*

$$\|x - y\| \leq \|x - g(y)\| \quad \text{for all } x \in C. \tag{27}$$

(S) *Let $D \subset C$ be a finite δ-grid of C for some $\delta > 0$, and let $g : D \to \mathbb{R}^m$ be a function. Then there exists a probability distribution η on D that is of type MM and has support of size at most $m + 3$, such that*

$$\mathbb{E}_{y\sim\eta}\left[\|x - y\|^2\right] \leq \mathbb{E}_{y\sim\eta}\left[\|x - g(y)\|^2\right] + \delta^2 \quad \text{for all } x \in C. \tag{28}$$

(AD) *Let $D \subset C$ be a finite δ-grid of C for some $\delta > 0$, and let $g : D \to \mathbb{R}^m$ be a function. Then there exists a probability distribution η on D that is δ-local, of type FP, has support of size at most $m + 1$, and satisfies (28).*

Proof. We will use the following easy-to-verify identity

$$\|x - y\|^2 - \|x - z\|^2 = 2(z - y) \cdot (x - y) - \|z - y\|^2 \tag{29}$$

[54] The formulations here are seemingly slightly weaker than those of Foster and Hart (2021), but they are still equivalent to Brouwer's fixed point theorem and von Neumann's minimax theorem, respectively; see Remarks (c) and (d) below.

(with $z = g(y)$) to go from the statements in Foster and Hart (2021) to the present ones.

(D) The Fixed Point Outgoing Theorem 4 of Foster and Hart (2021) applied to the function $f(x) = g(x) - x$ yields a point $y \in C$ such that for all $x \in C$ we have

$$(g(y) - y) \cdot (x - y) \leq 0,$$

and thus $\|x - y\|^2 - \|x - g(y)\|^2 \leq 0$, by (29) with $z = g(y)$.

(S) The Minimax Outgoing Theorem 5 of Foster and Hart (2021) applied to the function $f(x) = g(x) - x$ yields a distribution $\eta \in \Delta(D)$ such that for all $x \in C$ we have

$$\mathbb{E}_{y\sim\eta}\left[(g(y) - y) \cdot (x - y)\right] \leq \delta\, \mathbb{E}_{y\sim\eta}\left[\|g(y) - y\|\right], \tag{30}$$

and thus, by (29) with $z = g(y)$,

$$\mathbb{E}_{y\sim\eta}\left[\|x - y\|^2 - \|x - g(y)\|^2\right] \leq \mathbb{E}_{y\sim\eta}\left[2\delta \|g(y) - y\|\right] - \mathbb{E}_{y\sim\eta}\left[\|g(y) - y\|^2\right],$$

which gives (28) since $2\delta \|g(y) - y\| \leq \delta^2 + \|g(y) - y\|^2$.

(AD) The same proof as for (S), but now using the Almost Deterministic Outgoing Theorem 7 of Foster and Hart (2021). □

What (27) says is that y is closer than $g(y)$ to each point x in C; similarly, (28) says that the random y with distribution η is closer on average than $g(y)$ (within a δ-tolerance) to each point x in C. To get some intuition, let $\lambda \equiv \lambda(x) := \|x - y\|^2 - \|x - g(y)\|^2$; if $g : C \to C$ (as in Brouwer's 1912 fixed point theorem) then condition (27), which says that $\lambda \leq 0$ for every $x \in C$, is equivalent to $g(y) = y$, i.e., to y being a fixed point of g (indeed, for a fixed point y we have $\lambda = 0$ for all x; conversely, for $x = g(y)$, which is a point in C, we get $\lambda = \|g(y) - y\|^2 \leq 0$, and thus $g(y) = y$). Condition (28) extends this by requiring that $\lambda \leq 0$ hold approximately *on average*, i.e., $\mathbb{E}[\lambda] \leq \delta^2$, for every $x \in C$. This suggests (28) as a suitable concept of a "stochastic approximate fixed point" (note that a point y such that y and $g(y)$ are close—a natural attempt to define an approximate fixed point concept—need not exist in general: take, for example, the function $g : [0, 1] \to [0, 1]$ given by $g(x) = 1$ for $x \leq 1/2$, and $g(x) = 0$ for $x > 1/2$, for which $|g(x) - x| \geq 1/2$ for all x).

Remarks. *(a)* It suffices to consider functions g whose range is included in[55] C. Indeed, replacing g with the function $\hat{g}$ given by $\hat{g}(x) := \mathrm{proj}_C g(x)$ for every x (which is continuous when g is continuous) can only decrease the right-hand sides of inequalities (27) and (28) (because $\|x - \mathrm{proj}_C z\| \leq \|x - z\|$ for every $x \in C$ and $z \in \mathbb{R}^m$), and so if they hold for $\hat{g}$ then they hold for g as well. For a direct proof of Theorem 11 (D) by Brouwer's (1912)

[55] In the application of these results to calibration (both in Foster and Hart 2021 and in the present paper) the functions are always into C (for instance, $g(c) = \bar{a}_{t-1}(c)$).

fixed point theorem, let y be a fixed point of $\hat{g}$; then $\|x-y\| = \|x-\hat{g}(y)\| \leq \|x-g(y)\|$ for all $x \in C$.

(b) A direct proof of Theorem 11 (S) by von Neumann's (1928) minimax theorem for finite games is as follows. Let $\delta_0 := \max_{x \in C} \operatorname{dist}(x, D) < \delta$ (see the proof of Theorem 5 in Foster and Hart 2021), put $\delta_1 := (\delta^2 - \delta_0^2)/4\gamma > 0$, and take $D_1 \subset C$ to be a finite δ_1-grid of C. Consider the finite two-person zero-sum game where the maximizer chooses $x \in D_1$, the minimizer chooses $y \in D$, and the payoff is $\|x-y\|^2 - \|x-g(y)\|^2$, where $g : D \to C$ is the given function (use Remark (a) above). For every mixed strategy $\xi \in \Delta(D_1)$ of the maximizer, let $\bar{x} := \mathbb{E}_{x \sim \xi}[x] \in C$ be its expectation; the minimizer can make the payoff $\leq \delta_0^2$ by choosing a point y on the grid D that is within δ_0 of $\bar{x}$, since[56]

$$\begin{aligned} \mathbb{E}_{x \sim \xi}\left[\|x-y\|^2\right] &= \mathbb{E}_{x \sim \xi}\left[\|x-\bar{x}\|^2\right] + \|\bar{x}-y\|^2 \leq \mathbb{E}_{x \sim \xi}\left[\|x-\bar{x}\|^2\right] + \delta_0^2 \\ \mathbb{E}_{x \sim \xi}\left[\|x-g(y)\|^2\right] &= \mathbb{E}_{x \sim \xi}\left[\|x-\bar{x}\|^2\right] + \|\bar{x}-g(y)\|^2 \geq \mathbb{E}_{x \sim \xi}\left[\|x-\bar{x}\|^2\right]. \end{aligned}$$

Therefore, by the minimax theorem, the minimizer can guarantee that the payoff is $\leq \delta_0^2$; i.e., there is a mixed strategy $\eta \in \Delta(D)$ such that

$$\mathbb{E}_{y \sim \eta}\left[\|x-y\|^2 - \|x-g(y)\|^2\right] \leq \delta_0^2 \tag{31}$$

for every $x \in D_1$. Now for every $x \in C$ there is $x' \in D_1$ with $\|x-x'\| < \delta_1$, and so $\left|\|x-z\|^2 - \|x'-z\|^2\right| < 2\gamma\delta_1$ for any z in C; adding this inequality for $z = y$ and for $z = g(y)$ to (31) yields, by the definition of δ_1, the claimed result (28).

(c) Theorem 11 (D) is equivalent to Brouwer's (1912) fixed point theorem. Indeed, the former has been proved by using the latter (see Foster and Hart 2021 or Remark (a) above); conversely, given a continuous function $g : C \to C$, inequality (27) for the point $x = g(y)$ (which is in C) yields $g(y) = y$.

(d) Theorem 11 (S) is equivalent to von Neumann's (1928) minimax theorem for finite games. Indeed, the former has been proved by using the latter (see Foster and Hart 2021 or Remark (b) above); conversely, we will show that the former yields Corollary 6 of Foster and Hart (2021), from which the minimax theorem follows by Remark 4 to Corollary 6 in Appendix A3.3 of Foster and Hart (2021). For this Corollary 6, let $f : C \to \mathbb{R}^m$ be a bounded function, say $\|f(x)\| \leq M$ for all $x \in C$, and let $\varepsilon > 0$. Applying (28) to the function $g(x) = x + \delta f(x)$ and a finite δ-grid D of C yields a distribution $\eta \in \Delta(C)$ such that $\mathbb{E}\left[2\delta f(y) \cdot (x-y)\right] \leq \mathbb{E}\left[\delta^2 \|f(y)\|^2\right] + \delta^2 \leq \delta^2(M^2+1)$ (use the identity (29)), and thus, by choosing δ so that $\delta(M^2+1)/2 \leq \varepsilon$, we get $\mathbb{E}\left[f(y) \cdot (x-y)\right] \leq \varepsilon$, which is the result of Corollary 6 of Foster and Hart (2021).

(e) As in Remark 1 to Theorem 5 in Appendix A3.2 of Foster and Hart (2021), one may use a limit argument (which, however, no longer yields the distribution η by

[56] This argument holds for any distribution ξ on the whole set C (not just on D_1).

a finite minimax construct) to replace δ in (30), and thus in (28), with $\delta_0 \equiv \delta_0(D) := \max_{x\in C} \operatorname{dist}(x, D) < \delta$. Thus,

$$\begin{aligned} & \inf_{\eta\in\Delta(D)} \sup_{x\in C} \mathbb{E}_{y\sim\eta} \left[\|x-y\|^2 - \|x-g(y)\|^2\right] \\ = & \sup_{\xi\in\Delta(C)} \inf_{y\in D} \mathbb{E}_{x\sim\xi} \left[\|x-y\|^2 - \|x-g(y)\|^2\right] \le \delta_0^2. \end{aligned}$$

Moreover, the δ_0^2 bound is tight (i.e., it cannot be replaced by any smaller constant): let $x_0 \in C$ be such that $\operatorname{dist}(x_0, D) = \delta_0$; then, for the constant function $g \equiv x_0$ (i.e., $g(y) = x_0$ for all $y \in D$), we have $\|x_0 - y\| \ge \delta_0$ and $\|x_0 - g(y)\| = 0$ for all $y \in D$.

A.4 Refined Refinement

In this appendix we prove formally that the refinement score is monotonically decreasing with respect to refining the binning; this yields in particular $\mathcal{R}_t^{\mathbf{b}^1,...,\mathbf{b}^N} \le \mathcal{R}_t^{\mathbf{b}^n}$ for each $n = 1, ..., N$ (Section 8) and also $\mathcal{R}_t^{\mathbf{b},\Pi} \le \mathcal{R}_t^{\mathbf{b}}$ and $\mathcal{R}_t^{\mathbf{b},\Pi} \le \mathcal{R}_t^{\Pi}$ (Section 7).

We consider general fractional binnings. Let I be a finite or countably infinite collection of bins, and consider a sequence $(z_s)_{s\ge1}$ (namely, $z_s = a_s - c_s$) such that at time s the fraction $\lambda_s(i) \ge 0$ of z_s is assigned to bin i for each $i \in I$, where $\sum_{i\in I} \lambda_s(i) = 1$ (the specific way in which these weights are determined will not matter). Fix the horizon $t \ge 1$ (we will thus drop the subscript t); the refinement score is

$$\mathcal{R} = \frac{1}{t} \sum_{i\in I} \sum_{s=1}^{t} \lambda_s(i)(z_s - \bar{z}(i))^2,$$

where, for each i in I,

$$\bar{z}(i) = \frac{\sum_{s=1}^{t} \lambda_s(i) z_s}{\sum_{s=1}^{t} \lambda_s(i)}$$

is the average of bin[57] i.

As in Section 2.1, let the two-dimensional random variable (Z, U) take the value (z_s, i) with probability $\lambda_s(i)/t$ for each $s = 1, ..., t$ and $i \in I$ (note that $\sum_{s\le t} \sum_{i\in I} \lambda_s(i)/t = 1$); thus, $\mathbb{P}[(Z, U) = (z, i)] = (1/t) \sum_{s\le t: z_s = z} \lambda_s(i)$, which is the average, over all periods

[57] Provided that $\sum_{s\le t} \lambda_s(i) > 0$.

$s = 1, ..., t$, of the probability that the value z goes into bin i. We then have

$$\begin{aligned}
\mathbb{P}[U = i] &= \sum_{s=1}^{t} \frac{\lambda_s(i)}{t}, \\
\mathbb{E}[Z|U = i] &= \frac{1}{\mathbb{P}[U = i]} \sum_{s=1}^{t} \left(\frac{\lambda_s(i)}{t}\right) z_s \;=\; \bar{z}(i), \\
\mathbb{V}ar[Z|U = i] &= \frac{1}{\mathbb{P}[U = i]} \sum_{s=1}^{t} \left(\frac{\lambda_s(i)}{t}\right) (z_s - \bar{z}(i))^2, \text{ and} \\
\mathbb{E}[\mathbb{V}ar[Z|U]] &= \sum_{i \in I} \mathbb{P}[U = i]\ \mathbb{V}ar[Z|U = i] \;=\; \mathcal{R}. \qquad (32)
\end{aligned}$$

Now assume that we are given another collection of bins J together with binning weights $\mu_s(j) \geq 0$, where $\sum_{j \in J} \mu_s(j) = 1$ for each s. The J-binning is a *coarsening* of the I-binning (equivalently, the I-binning is a *refinement* of the J-binning) if there is a function $\phi : I \to J$ such that $\mu_s(j) = \sum_{i:\phi(i)=j} \lambda_s(i)$; that is, for each j in J the j-bin is the union of the set $\phi^{-1}(j) = \{i \in I : \phi(i) = j\}$ of i-bins in I. Letting U_I and U_J be the random variables corresponding to the I-binning and the J-binning, respectively, we have $U_J = \phi(U_I)$, because being assigned to an i-bin for $i \in I$ translates to being assigned to the j-bin for $j = \phi(i) \in J$. Let $\mathcal{R}_I$ and $\mathcal{R}_J$ be the refinement scores corresponding to the I-binning and the J-binning, respectively.

Proposition 12 *If the J-binning is a coarsening of the I-binning then*

$$\mathcal{R}_J = \mathbb{E}[\mathbb{V}ar[Z|U_J]] \geq \mathbb{E}[\mathbb{V}ar[Z|U_I]] = \mathcal{R}_I.$$

Proof. Let $\mathcal{F}_1, \mathcal{F}_2$ be two σ-fields such that $\mathcal{F}_1 \subseteq \mathcal{F}_2$, i.e., $\mathcal{F}_1$ is a coarsening of $\mathcal{F}_2$, and let Z be a random variable. We will show that

$$\mathbb{E}[\mathbb{V}ar[Z|\mathcal{F}_1]] \geq \mathbb{E}[\mathbb{V}ar[Z|\mathcal{F}_2]], \qquad (33)$$

which yields the result by (32).

Applying the classic inequality $\mathbb{V}ar[X] = \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq \mathbb{E}\left[\|X - x\|^2\right]$ for any random variable X and any constant x (i.e., the expected square deviation from a constant is minimized when the constant equals the expectation) to $Z|\mathcal{F}_2$ we get (a.s.)

$$\mathbb{V}ar[Z|\mathcal{F}_2] \leq \mathbb{E}\left[\|Z - \mathbb{E}[Z|\mathcal{F}_1]\|^2 |\mathcal{F}_2\right],$$

because $\mathbb{E}[Z|\mathcal{F}_1]$ is constant given $\mathcal{F}_2$ (since $\mathcal{F}_1$ is a coarsening of $\mathcal{F}_2$). Taking expectation conditional on $\mathcal{F}_1$ yields on the right-hand side $\mathbb{E}\left[\|Z - \mathbb{E}[Z|\mathcal{F}_1]\|^2 |\mathcal{F}_1\right]$ (again, by $\mathcal{F}_1 \subseteq$

$\mathcal{F}_2$), which is the conditional variance $\mathbb{V}ar\left[Z|\mathcal{F}_1\right]$, and so we have (a.s.)

$$\mathbb{E}\left[\mathbb{V}ar\left[Z|\mathcal{F}_2\right]|\mathcal{F}_1\right] \leq \mathbb{V}ar\left[Z|\mathcal{F}_1\right].$$

Taking overall expectation yields (33), and thus completes the proof. □

Applying Proposition 12 with ϕ being a projection, such as $\phi(b^1, ..., b^N) = b^n$, yields the needed inequalities.

A.5 General Brier Score Decomposition

We show here that the decomposition of the Brier score (1) into the refinement and the calibration scores holds for any fractional binning $\Pi = (w_i)_{i=1}^I$, i.e.,

$$\mathcal{B}_t = \mathcal{R}_t^\Pi + \mathcal{K}_t^\Pi.$$

In the notation of the previous Section A.4, this is

$$\mathbb{E}\left[\|Z\|^2\right] = \mathbb{E}\left[\mathbb{E}[\|Z\|^2|U]\right] = \mathbb{E}\left[\mathbb{V}ar\left[Z|U\right]\right] + \mathbb{E}\left[\|\mathbb{E}\left[Z|U\right]\|^2\right],$$

which follows from applying the identity $\mathbb{E}\left[X^2\right] = \mathbb{V}ar\left[X\right] + \mathbb{E}\left[X\right]^2$ to each one of the m coordinates of $Z|U$, summing up, and then taking overall expectation.

A.6 Continuous Calibration

We first generalize Proposition 2 to weighted variances. Let $(x_n)_{n\geq 1}$ be a sequence of vectors in a Euclidean space (or, more generally, in a normed vector space), let $(\lambda_n)_{n\geq 1}$ be a sequence of weights in $[0, 1]$, and put $\Lambda_n := \sum_{i=1}^n \lambda_i$.

Proposition 13 *For every $n \geq 1$ we have*[58]

$$\frac{1}{\Lambda_n}\sum_{i=1}^n \lambda_i \|x_i - \bar{x}_n\|^2 = \frac{1}{\Lambda_n}\sum_{i=1}^n \lambda_i \left(1 - \frac{\lambda_i}{\Lambda_i}\right)\|x_i - \bar{x}_{i-1}\|^2,$$

where $\bar{x}_m := (1/\Lambda_m)\sum_{i=1}^m \lambda_i x_i$ denotes the weighted average of[59] *$x_1, ..., x_m$.*

Proof. Put $s_n := \sum_{i=1}^n \lambda_i \|x_i - \bar{x}_n\|^2$; we claim that

$$s_n = s_{n-1} + \lambda_n \left(1 - \frac{\lambda_n}{\Lambda_n}\right)\|x_n - \bar{x}_{n-1}\|^2. \tag{34}$$

[58] When $\Lambda_n = 0$, and thus $\lambda_i = 0$ for all $i - 1, ..., n$, we put $(1/\Lambda_n)\sum_{i=1}^n \lambda_i z_i = 0$.

[59] The sum on the right-hand side effectively starts from $i = 2$, and so, again, it does not matter how $\bar{x}_0$ is defined.

Indeed, assume that $\Lambda_n > 0$ (otherwise both sides vanish) and $\bar{x}_{n-1} = 0$ (without loss of generality, since subtracting a constant from all the x_i does not affect any term); then $\bar{x}_n = (\lambda_n/\Lambda_n)x_n$, and so, using $s_n = \sum_{i=1}^n \lambda_i ||x_i||^2 - \Lambda_n ||\bar{x}_n||^2$, we get

$$s_n - s_{n-1} = \left(\sum_{i=1}^{n} \lambda_i \left\| x_i \right\|^2 - \Lambda_n \left\| \frac{\lambda_n}{\Lambda_n} x_n \right\|^2 \right) - \sum_{i=1}^{n-1} \lambda_i \left\| x_i \right\|^2 = \lambda_n \left\| x_n \right\|^2 - \frac{\lambda_n^2}{\Lambda_n} \left\| x_n \right\|^2,$$

which is precisely $\lambda_n \left(1 - \lambda_n/\Lambda_n\right) \left\| x_n - \bar{x}_{n-1} \right\|^2$.

Applying (34) recursively yields the result. □

Let $v_n := (1/\Lambda_n) \sum_{i=1}^n \lambda_i \left\| x_i - \bar{x}_n \right\|^2$ denote the weighted variance of $x_1, ..., x_n$, and put $\widetilde{v}_n := (1/\Lambda_n) \sum_{i=1}^n \lambda_i \left\| x_i - \bar{x}_{i-1} \right\|^2$ for the corresponding *online weighted variance* of[60] $x_1, ..., x_n$. Proposition 13 gives $\widetilde{v}_n - v_n = (1/\Lambda_n) \sum_{i=1}^n (\lambda_i^2/\Lambda_i) \left\| x_i - \bar{x}_{i-1} \right\|^2$, and so, by inequality (22) in Foster and Hart (2021),

$$0 \leq \widetilde{v}_n - v_n \leq \xi^2 \frac{1}{\Lambda_n} \sum_{i=1}^{n} \frac{\lambda_i^2}{\Lambda_i} \leq \xi^2 \frac{\ln \Lambda_n + 2}{\Lambda_n}, \tag{35}$$

where $\xi := \max_{1 \leq i,j \leq n} \left\| x_i - x_j \right\|$.

Second, we prove Proposition 7, which shows that the online refinement score $\widetilde{\mathcal{R}}_t^{\mathbf{b},\Pi}$ is close to the (offline) refinement score.

Proof of Proposition 7. We have $\widetilde{\mathcal{R}}_t^{\mathbf{b},\Pi} - \mathcal{R}_t^{\mathbf{b},\Pi} = (1/t) \sum_{b \in B} \sum_{i \in I} \mu_t(b,i)$, where

$$\mu_t(b,i) := \sum_{s=1}^{t} \lambda_s(b,i) \left\| a_s - c_s - e_{s-1}(b,i) \right\|^2 - \sum_{s=1}^{t} \lambda_s(b,i) \left\| a_s - c_s - e_t(b,i) \right\|^2$$

for each $(b,i) \in B \times I$. Proposition 13, specifically, (35), yields

$$0 \leq \mu_t(b,i) \leq 4\gamma^2 (\ln n_t(b,i) + 2) \leq 4\gamma^2 (\ln t + 2) \tag{36}$$

(because $\left\| a - c - e \right\| \leq 2\gamma$—since $\left\| a - c \right\| \leq \gamma$ and so $\left\| e \right\| \leq \gamma$—and $n_t(b,i) \leq t$). For each finite $J \subseteq I$, summing over all (b,i) in $B \times J$ yields

$$0 \leq \frac{1}{t} \sum_{b \in B} \sum_{i \in J} \mu_t(b,i) \leq 4\gamma^2 |B| \ |J| \ \frac{\ln t + 2}{t}. \tag{37}$$

When I is finite we are thus done. When I is infinite, for every $\varepsilon > 0$ there is a finite $J \subset I$ such that $\sum_{i \in I \setminus J} w_i(c) \leq \varepsilon$ for all $c \in C$; such a finite J exists by Dini's theorem

[60] Again, take $\bar{x}_0$ to be an arbitrary element of the convex hull of the x_i.

(see (9) in Foster and Hart 2021). For $i \in I \backslash J$ we get

$$\begin{aligned} 0 &\leq \frac{1}{t}\sum_{b\in B}\sum_{i\in I\backslash J}\mu_t(b,i) \leq \frac{1}{t}\sum_{b\in B}\sum_{i\in I\backslash J}\sum_{s=1}^{t}\lambda_s(b,i)\left\|a_s - c_s - e_{s-1}(b,i)\right\|^2 \\ &\leq 4\gamma^2\frac{1}{t}\sum_{s=1}^{t}\sum_{i\in I\backslash J}\lambda_s(b_s,i) \leq 4\gamma^2\frac{1}{t}\sum_{s=1}^{t}\varepsilon = 4\gamma^2\varepsilon. \end{aligned}$$

Adding this to (37) yields

$$0 \leq \widetilde{\mathcal{R}}_t^{\mathbf{b},\Pi} - \mathcal{R}_t^{\mathbf{b},\Pi} \leq 4\gamma^2|B|\,|J|\,\frac{\ln t + 2}{t} + 4\gamma^2\varepsilon,$$

which is less than, say, $5\gamma^2\varepsilon$ for all large enough t. The result follows since ε was arbitrary; moreover, all the above inequalities are uniform over all sequences $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. □

A.7 Multi-Calibeating: Improved Error Terms

The multi-calibeating procedure of Section 8 yields an error that is proportional to the product of the sizes of the sets B^n, i.e., $\prod_{n=1}^{N}|B^n|$, which increases exponentially with N. We provide in this appendix two approaches that yield better errors terms: $O(\sqrt{N}/\sqrt{t})$ by a Blackwell approachability construct in Section A.7.1 (see (40)), and $O((\max_n|B^n|+N)\log t/t)$ by an online linear regression construct in Section A.7.2 (see (43)).[61]

We will use throughout the superscript n instead of the more cumbersome $\mathbf{b}^n$, e.g., $\mathcal{R}_t^n$ for $\mathcal{R}_t^{\mathbf{b}^n}$, and $\bar{a}_{t-1}^n(b_t^n)$ for $\bar{a}_{t-1}^{\mathbf{b}^n}(b_t^n)$.

A.7.1 A Blackwell Approachability Approach

We use here a construct along the lines of the vector approachability of Blackwell (1950).[62]

Put $x_t^n := \|a_t - c_t\|^2 - \left\|a_t - \bar{a}_{t-1}^n(b_t^n)\right\|^2$; then

$$\bar{x}_t^n := \frac{1}{t}\sum_{s=1}^{t}x_s^n = \mathcal{B}_t^{\mathbf{c}} - \widetilde{\mathcal{R}}_t^n. \tag{38}$$

Let $x_t := (x_t^n)_{n=1,\ldots,N} \in \mathbb{R}^N$ and $\bar{x}_t := (\bar{x}_t^i)_{n=1,\ldots,N} \in \mathbb{R}^N$ be the corresponding N-dimensional vectors; given the history $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{b}_t)$, the vector x_t is determined by c_t and a_t. We will show that the negative orthant[63] $\mathbb{R}_-^N$ of $\mathbb{R}^N$ is approachable by the c-player;

[61] For large t, an error of an order of magnitude of $1/\sqrt{t}$ is of course worse than $\log t/t$. However, for small t the former may well be smaller than the latter.

[62] With continuous actions taking the place of mixed actions (which dispenses from the use of probabilities and laws of large numbers).

[63] Some notation: $\mathbb{R}_+^N = \{x \in \mathbb{R}^N : x \geq 0\}$ and $\mathbb{R}_-^N = \{x \in \mathbb{R}^N : x \leq 0\}$; for real $x \in \mathbb{R}$, put $[x]_+ = \max\{x, 0\}$ and $[x]_- = \min\{x, 0\}$; for a vector $x \in \mathbb{R}^N$, put $[x]_+ = ([x_1]_+, ..., [x_N]_+)$, and $[x]_- = ([x_1]_-, ..., [x_N]_-)$.

i.e., that there is a **b**-based forecasting procedure ζ such that $\sup_{\mathbf{a}_t,\mathbf{b}_t} \mathrm{dist}(\bar{x}_t, \mathbb{R}^N_-) \to 0$ as $t \to \infty$.

To this end we claim that for every $\lambda \in \mathbb{R}^N_+$ there is $c_t \in C$ such that

$$\lambda \cdot x_t \leq 0 \text{ for every } a_t \in A \tag{39}$$

(this is the Blackwell condition here). This of course holds when $\lambda = 0$; otherwise, assuming without loss of generality that $\sum_{n=1}^N \lambda_n = 1$ (rescale λ as needed), we have

$$\begin{aligned}
\lambda \cdot x_t &= \sum_{n=1}^N \lambda_n \left(\|a_t - c_t\|^2 - \left\|a_t - \bar{a}^n_{t-1}(b^n_t)\right\|^2 \right) \\
&= \|a_t - c_t\|^2 - \sum_{n=1}^N \lambda_n \left\|a_t - \bar{a}^n_{t-1}(b^n_t)\right\|^2 \\
&\leq \|a_t - c_t\|^2 - \left\| a_t - \sum_{n=1}^N \lambda_n \bar{a}^n_{t-1}(b^n_t) \right\|^2
\end{aligned}$$

(the inequality is by the convexity of $\|\cdot\|^2$), and so, by taking $c_t = \sum_{n=1}^N \lambda_n \bar{a}^n_{t-1}(b^n_t)$, one guarantees that the final expression vanishes, and thus $\lambda \cdot x_t \leq 0$, for any a_t.

Let ζ be the procedure whereby at time t one chooses $c_t \in C$ so as to guarantee $[\bar{x}_{t-1}]_+ \cdot x_t \leq 0$ for all $a_t \in A$ (i.e., condition (39) for $\lambda = [\bar{x}_{t-1}]_+$); thus, c_t is arbitrary when $\bar{x}_{t-1} \leq 0$ (i.e., $[\bar{x}_{t-1}]_+ = 0$), and is otherwise given by

$$c_t = \frac{\sum_{n=1}^N \left[\bar{x}^n_{t-1}\right]_+ \bar{a}^n_{t-1}(b^n_t)}{\sum_{n=1}^N \left[\bar{x}^n_{t-1}\right]_+}.$$

Putting $X_t := t\bar{x}_t$, we have

$$\begin{aligned}
\mathrm{dist}^2(X_t, \mathbb{R}^N_-) &\leq \|(x_t + X_{t-1}) - [X_{t-1}]_-\|^2 = \|x_t + [X_{t-1}]_+\|^2 \\
&= \|[X_{t-1}]_+\|^2 + 2[X_{t-1}]_+ \cdot x_t + \|x_t\|^2 \\
&\leq \mathrm{dist}^2(X_{t-1}, \mathbb{R}^N_-) + \gamma^4 N,
\end{aligned}$$

where the first inequality is by $\mathrm{dist}(X_t, \mathbb{R}^N_-) \leq \|X_t - [X_{t-1}]_-\|$ (since $[X_{t-1}]_- \in \mathbb{R}^N_-$), and the second inequality is by the choice of c_t (since $[X_{t-1}]_+ = (t-1)[\bar{x}_{t-1}]_+$) for the middle term, and $|x^n_t| \leq \gamma^2$ (since $a_t, c_t, \bar{a}^n_{t-1}(b_t)$ are all in C) for all n for the third term. Applying this recursively yields

$$t^2 \mathrm{dist}^2(\bar{x}_t, \mathbb{R}^N_-) = \mathrm{dist}^2(X_t, \mathbb{R}^N_-) \leq (\gamma^4 N)t,$$

and so $\text{dist}^2(\bar{x}_t, \mathbb{R}^N_-) \leq (\gamma^4 N)/t$, which gives

$$\max_{1\leq n\leq N} \bar{x}^n_t \leq \max_{1\leq n\leq N} [\bar{x}^n_t]_+ \leq \left\|[\bar{x}_t]_+\right\| = \text{dist}(\bar{x}_t, \mathbb{R}^N_-) \leq \gamma^2\sqrt{N}\frac{1}{\sqrt{t}}.$$

By (38) and Proposition 1 we get

$$\max_{1\leq n\leq N} (\mathcal{B}^{\mathfrak{c}}_t - \mathcal{R}^n_t) \leq \gamma^2\sqrt{N}\frac{1}{\sqrt{t}} + \gamma^2|B^n|\frac{\ln t + 1}{t}, \tag{40}$$

which is $\sim \gamma^2\sqrt{N}/\sqrt{t}$ as $t \to \infty$.

A.7.2 An Online Linear Regression Approach

Assume that $C \subseteq [-\gamma_0, \gamma_0]^m$ (see the remark below on the relation between γ and γ_0), and put $c_s = (c_{i,s})_{i=1,...,m} \in C$ and $a_s = (a_{i,s})_{i=1,...,m.} \in A$. For each $t \geq 1$ let $x^n_{i,t} := \bar{a}^n_{i,t-1}(b^n_t)$ (this is the average of the ith coordinates of a_s over all periods $s \leq t$ in which $b^n_s = b^n_t$), and put $x_{i,t} := (x^n_{i,t})_{n=1,...,N} \in \mathbb{R}^N$.

For each coordinate $i = 1, ..., m$, consider the linear regression problem, regularized by adding the strictly convex term $\alpha \|\theta\|^2$ for some $\alpha > 0$, of minimizing

$$\mathcal{L}_{i,t}(\theta) := \frac{1}{t}\left(\sum_{s=1}^{t}(a_{i,s} - \theta \cdot x_{i,s})^2 + \alpha \|\theta\|^2\right)$$

over $\theta \in \mathbb{R}^N$; let $\mathcal{L}^*_{i,t}$ denote this minimum. For each $n = 1, ..., N$, when θ equals the nth unit vector $e^n \in \mathbb{R}^N$, we have

$$\mathcal{L}_{i,t}(e^n) = \frac{1}{t}\sum_{s\leq t}(a_{i,s} - \bar{a}^n_{i,s-1}(b^n_s))^2 + \frac{\alpha}{t};$$

summing over $i = 1, ..., m$ we get

$$\sum_{i=1}^{m}\mathcal{L}^*_{i,t} \leq \sum_{i=1}^{m}\mathcal{L}_{i,t}(e^n) \leq \widetilde{\mathcal{R}}^n_t + \frac{m\alpha}{t}. \tag{41}$$

The "forward algorithm" of Azoury and Warmuth (2001) applied to each coordinate i separately yields an online procedure that generates at each time t a vector $\theta_{i,t} \in \mathbb{R}^N$ (that depends on the history $a_{i,1}, ..., a_{i,t-1}$ and $x_{i,1}, ..., x_{i,t-1}$ as well as on $x_{i,t}$) such that

$$\sum_{s=1}^{t}(a_{i,s} - \theta_{i,s} \cdot x_{i,s})^2 \leq t\mathcal{L}^*_{i,t} + \gamma_0 N \ln\left(\frac{\gamma_0}{\alpha}t + 1\right) \tag{42}$$

is guaranteed for any sequence[64] $(a_{i,s})_{s\geq 1}$.

[64]This is Theorem 5.6 of Azoury and Warmuth (2001); in the notation there, $X = \max_{n,s}\left|x^n_{i,s}\right| \leq \gamma_0$

Combining these m algorithms yields an online **b**-based procedure (because x_t is determined by b_t and the history); the vectors $\theta_{i,s}$ for $i = 1, ..., m$ together yield a point $\hat{c}_s := (\theta_{i,s} \cdot x_{i,s})_{i=1,...,m} \in \mathbb{R}^m$. Let $c_s := \mathrm{proj}_C(\hat{c}_s)$ be the closest point to $\hat{c}_s$ in C (it is well defined since C is a nonempty convex compact set); then any point in C, in particular a_s, is closer to c_s than to $\hat{c}_s$, which yields

$$\|a_s - c_s\|^2 \le \|a_s - \hat{c}_s\|^2 = \sum_{i=1}^{m} (a_{i,s} - \theta_{i,s} \cdot x_{i,s})^2.$$

Averaging over $s \le t$ gives

$$\begin{aligned} \mathcal{B}_t^{\mathfrak{c}} &= \frac{1}{t} \sum_{s=1}^{t} \|a_s - c_s\|^2 \le \frac{1}{t} \sum_{i=1}^{m} \sum_{s=1}^{t} (a_{i,s} - \theta_{i,s} \cdot x_{i,s})^2 \\ &\le \sum_{i=1}^{m} \mathcal{L}_{i,t}^* + \frac{m\gamma_0 N}{t} \ln\left(\frac{\gamma_0}{\alpha} t + 1\right) \end{aligned}$$

by (42). Recalling (41) and Proposition 1 yields

$$\mathcal{B}_t^{\mathfrak{c}} - \mathcal{R}_t^n \le \frac{m\gamma_0 N}{t} \ln\left(\frac{\gamma_0}{\alpha} t + 1\right) + \frac{m\alpha}{t} + \gamma^2 |B^n| \frac{\ln t + 1}{t} = O\left(\frac{\log t}{t}\right). \tag{43}$$

Remarks. *(a)* A connection between γ_0 and γ is as follows. The set $C \subset \mathbb{R}^m$, whose diameter is γ, can be enclosed in a ball of radius r, where $\gamma/2 \le r \le \gamma\sqrt{m/(2m+2)}$ by Jung's (1901) theorem. Since translating the set C does not matter (only differences $a - c$ do), we can assume without loss of generality that $C \subseteq \bar{B}(0; r) \subseteq [-r, r]^m$, and so we can take $\gamma_0 = r$.

(b) For a fixed horizon t one may optimize α.

(c) The forecast c_t at time t of the above construction is given by the formula

$$c_{i,t} = \sum_{n=1}^{N} \theta_{i,t}^n \bar{a}_{t-1}^n (b_t^n),$$

where $\theta_{i,t}$ is the minimizer of $\mathcal{L}_{i,t}(\theta)$ as if we have $a_{i,t} = 0$ (the actual a_t is not known at this point); see Azoury and Warmuth (2001) for details and more explicit formulas.

(d) Since we use the inequalities $\mathcal{L}_{i,t}^* \le \mathcal{L}_{i,t}(\theta)$ only for θ equal to the unit vectors e^n in $\mathbb{R}^N$, it suffices to minimize $\mathcal{L}_{i,t}(\theta)$ over the convex hull of these vectors, that is, over the unit simplex $\Delta(N)$ of $\mathbb{R}^N$, as in Foster (1991).[65]

and $Y = \max_s |a_{i,s}| \le \gamma_0$. Note that there is a misprinted sign in the first line of the formula (5.17) there.

[65] Whose result would need to be generalized from the one-dimensional case of $A = \{0, 1\}$ and $C = [0, 1]$ to a general C. Note that, by Proposition 1, multi-calibeating is equivalent to being "as strong as," in terms of the Brier scores, each one of the N the sequences $(\bar{a}_{t-1}^n(b_t^n))_{t \ge 1}$, for $n = 1, ..., N$.

(e) An alternative approach is to first calibeat each forecaster separately (by Theorem 3) and then to combine these N calibeating forecasters (by a method such as Azoury and Warmuth's 2001).

References

Azoury, K. S. and M. K. Warmuth (2001), "Relative Loss Bounds for On-Line Density Estimation with the Exponential Family of Distributions," *Machine Learning* 43, 211–246.

Blackwell, D. (1956), "An Analog of the Minimax Theorem for Vector Payoffs," *Pacific Journal of Mathematics* 6, 1–8.

Brier, G. W. (1950), "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* 78, 1–3.

Brouwer, L. E. J. (1912), "Über Abbildung von Mannigfaltigkeiten," *Mathematische Annalen* 71, 97–115.

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

Dawid, A. (1982), "The Well-Calibrated Bayesian," *Journal of the American Statistical Association* 77, 605–613.

Forster, J. (1999), "On Relative Loss Bounds in Generalized Linear Regression," in *12th International Symposium on Fundamentals of Computation Theory (FCT '99)*, 269–280.

Foster, D. P. (1991), "Prediction in the Worst Case," *The Annals of Statistics* 19, 1084–1090.

Foster, D. P. (1999), "A Proof of Calibration via Blackwell's Approachability Theorem," *Games and Economic Behavior* 29, 73–78.

Foster, D. P. and S. Hart (2018), "Smooth Calibration, Leaky Forecasts, Finite Recall, and Nash Dynamics," *Games and Economic Behavior* 109, 271–293.

Foster, D. P. and S. Hart (2021), "Forecast Hedging and Calibration," *Journal of Political Economy* 129, 3447–3490. doi.org/10.1086/716559. `http://www.ma.huji.ac.il/hart/publ.html#calib-int`

Foster, D. P. and R. V. Vohra (1998), "Asymptotic Calibration," *Biometrika* 85, 379–390.

Hart, S. (1992), "Games in Extensive and Strategic Forms," in *Handbook of Game Theory, with Economic Applications*, R. J. Aumann and S. Hart (editors), North-Holland, Vol. 1, Chapter 2, 19–40.

Hart, S. (2021), "Calibrated Forecasts: The Minimax Proof," Center for Rationality DP-744, The Hebrew University of Jerusalem. `http://www.ma.huji.ac.il/hart/publ.html#calib-minmax`

Johnson, N. L., A. W. Kemp, and S. Kotz (2005), *Univariate Discrete Distributions*, 3rd edition, Wiley.

Jung, H. (1901), "Ueber die kleinste Kugel, die eine räumliche Figur einschliesst," *Journal für die Reine und Angewandte Mathematik* 123, 241–257.

Kuhn, H. W. (1953), "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games, Vol. II*, H. W. Kuhn and A. W. Tucker (editors), *Annals of Mathematics Studies* 28, Princeton University Press, 193–216.

Loève, M. (1978), *Probability Theory, Vol. II*, 4th edition, Springer.

Murphy, A. H. (1972), "Scalar and Vector Partitions of the Probability Score. Part I: Two-State Situation," *Journal of Applied Meteorology* 11, 273–282.

Oakes, D. (1985), "Self-Calibrating Priors Do Not Exist," *Journal of the American Statistical Association* 80, 339.

Olszewski, W. (2015), "Calibration and Expert Testing," in *Handbook of Game Theory, Vol. 4*, H. P. Young and S. Zamir (editors), Springer, 949–984.

Olszewski, W. and A. Sandroni (2008), "Manipulability of Future-Independent Tests," *Econometrica* 76, 1437–1466.

Sanders, F. (1963), "On Subjective Probability Forecasting," *Journal of Applied Meteorology* 2, 191–201.

Sandroni, A. (2003), "The Reproducible Properties of Correct Forecasts," *International Journal of Game Theory* 32, 151–159.

Shmaya, E. (2008), "Many Inspections are Manipulable," *Theorerical Economics* 3, 367–382.

Welford, B. P. (1962), "Note on a Method for Calculating Corrected Sums of Squares and Products," *Technometrics* 4, 419–420.

Vovk, V. (2001), "Competitive On-Line Statistics," *International Statistical Review* 69, 213–248.

von Neumann, J. (1928), "Zur Theorie der Gesellschaftsspiele, " *Mathematische Annalen* 100, 295–320.