"Calibeating": Beating Forecasters at Their Own Game*

Dean P. Foster† Sergiu Hart‡

December 7, 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.

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.* There is a large literature on calibration; see the

*Previous versions: February 2020; October 2021 (Hebrew University of Jerusalem, Center for Rationality DP-743), May 2022; October 2022 (arXiv:2209.04892v2; this is the long version of the preset paper, Foster and Hart 2022). We thank Drew Fudenberg, Benjy Weiss, the coeditor, and the referees for useful comments and suggestions. 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

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).

The *calibration score* $\mathcal{K}$ is defined as the average squared distance between forecasts and realized (relative) frequencies (i.e., the proportion of, say, rainy days), where each forecast is weighted by how often it has been used; evaluated after t days, this yields

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

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 in the days from 1 to t in which the forecast was p (giving weight $1/t$ to each day is the same as weighting each forecast by the proportion of days it has been used). 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. 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.)

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, specifically, $1/(4t^2)$, on odd days), 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:

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

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 the calibration score 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 high 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 $\bar{a}(c_s)$:

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

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 $\mathcal{R}$ (indeed, $\mathcal{R}$ is the "relabeling-minimum" Brier score; see Section 2.3.1), and changing the distribution of actions within each bin without changing their average does not affect $\mathcal{K}$. 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 (for perfect classification without calibration, 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).

Thus, our first conclusion is

> *Conclusion: Experts should better be tested by the Brier score and not by calibration 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 (which can only decrease $\mathcal{R}$; see Section 5 and Appendix A.4)—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. 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 always be 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. 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 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. 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.

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 minimax" 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—we obtain (Theorem 6 in Section 6 and Theorem 12 in Appendix A.6) *deterministic* calibeating procedures that are continuously calibrated. This requires the use of a fixed point tool, specifically, the "outgoing fixed point" result of Foster and Hart (2021); see Section III.D there, and Appendix A.2 here, for the important distinction between minimax and fixed point methods.

Next, we show that all the above results can be extended to simultaneously calibeating multiple forecasters (Theorem 7 in Section 7).

Finally, we comment on the use of the quadratic scores (such as $\|a-c\|^2$). This is standard in statistics (e.g., analysis of variance and linear regression), as it easily leads to useful decompositions, such as the Brier score being the sum of the refinement and calibration scores here. However, it raises the question of how much our results depend on the quadratic scores.[1] We believe that the ideas and approach here carry through for other scoring functions; in the long version of the paper (Foster and Hart 2022, Appendix A.9) we show this for another classic scoring rule, the logarithmic one.

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 should better be tested by the Brier score and not just by calibration, and show how to calibeat forecasts that are not calibrated: 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 $\mathrm{conv}(A)$ of A. Let $\gamma := \mathrm{diam}(C) \equiv \max_{c,c' \in C} \|c - c'\|$ 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 of 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 $\mathbf{c}_t = (c_s)_{1 \leq s \leq t}$.

[1] We thank the referee who posed this question.

2.1 The Calibration, Refinement, and Brier Scores

Fix a time horizon t. For each possible forecast x in C let[2]

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

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

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

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

and the *calibration score* is the average square calibration error, i.e.,[3]

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

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

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

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 $(K_t)^2 \le \mathcal{K}_t \le \gamma K_t$ (the first inequality by Jensen's inequality, the second by $\|e_t(x)\| \le \gamma$), and so $\mathcal{K}_t \to 0$ if and only if $K_t \to 0$.

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

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

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

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

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. This is a so-called "strictly proper scoring rule," which means that if the sequence $\mathbf{a}_t$ is generated by a probability distribution $\mathbb{P}$, then the unique minimizer of the expected Brier score is the sequence $c_s = \mathbb{P}[a_s|\mathbf{a}_{s-1}]$ of true conditional probabilities (assume for simplicity that C is the set of probability distributions $\Delta(A)$ on a finite set A).

One may assume for convenience[4] 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 we have

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

Using the identity $\mathbb{E}[X^2] = \mathbb{V}ar[X] + (\mathbb{E}[X])^2$ for each one of the m coordinates of $Z|U$, summing over 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). Appendix A.5 generalizes this to "fractional" binnings.

For each x the variance $v_t(x)$ of the x-bin is the minimum over $y \in C$ of $n_t(x)^{-1}\sum_{1\le s\le t:c_s=x} \|a_s - y\|^2$, which is attained when y equals the bin average $\bar{a}_t(x)$. Therefore, the refinement score is the Brier score where each bin label x is replaced by $\bar{a}_t(x)$, and this is the minimal Brier score over all relabelings of the bins:

$$\mathcal{R}_t = \min_{\phi} \mathcal{B}_t^{\phi(\mathbf{c})}, \tag{2}$$

[4]See Foster and Hart (2021); this matters also when generalizing to fractional binnings (Section A.6).

where the minimum is taken over all functions $\phi : C \to C$ (from current labels x to new labels y), and we write $\mathcal{B}_t^{\phi(\mathbf{c})}$ for the Brier score where the sequence $\mathbf{c}$ is replaced by[5] $\phi(\mathbf{c}) = (\phi(c_s))_{1 \le s \le t}$. Thus, starting from the Brier scoring rule, we could define the refinement score $\mathcal{R}$ as the "relabeling-minimum" Brier score, and the calibration score $\mathcal{K}$ as the "residual" score $\mathcal{B} - \mathcal{R}$.

2.2 Calibration

A stochastic *forecasting procedure* σ is a mapping $\sigma : \cup_{t \ge 1}(A^{t-1} \times C^{t-1}) \to \Delta(C)$; 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 $\varepsilon \ge 0$; a (stochastic) procedure σ is *ε-calibrated* (Foster and Vohra 1998) if[6,7]

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

(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, fixing the action sequence, $\mathbf{c}$ beats the Brier score of $\mathbf{b}$ by at least $\mathbf{b}$'s calibration score (i.e., $\mathcal{B}^{\mathbf{c}} \le \mathcal{B}^{\mathbf{b}} - \mathcal{K}^{\mathbf{b}}$ in the long run). Thus, if $\mathbf{b}$ is not calibrated, and hence its calibration score $\mathcal{K}^{\mathbf{b}}$ 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., to have 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; moreover, this should hold uniformly over all these sequences.

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,

[5] Joining two bins that have the same average does not affect the refinement score.

[6] 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.

[7] 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 Section 2.1, $(K_t)^2 \le \mathcal{K}_t$.

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 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[8] $\zeta : \cup_{t\geq 1}(A^{t-1} \times C^{t-1} \times B^t) \to \Delta(C)$. 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,

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

where the expectation $\mathbb{E}$ is over the random forecasts 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 (3) 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 (3) 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 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

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

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 (3) and (4) 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 the

[8] 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 $\mathbf{b}$).

notion of calibeating to be generalized 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

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

or, equivalently,

$$\overline{\lim_{t\to\infty}} \left(\sup_{\mathbf{a}_t\in A^t, \mathbf{b}_t\in B^t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \right) \leq \varepsilon^2. \tag{6}$$

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

Finally, calibeating can be formalized in terms of Brier scores only. Since the refinement score is the minimal Brier score over all relabelings of the bins (see (2)), it follows that calibeating amounts to getting the Brier score of $\mathbf{c}$ down to the "relabeling-minimum" Brier score of $\mathbf{b}$, i.e. (ignoring ε, $\overline{\lim}$, and sup), $\mathcal{B}_t^{\mathbf{c}} \leq \min_\phi \mathcal{B}_t^{\phi(\mathbf{b})}$, where the minimum is taken over all functions $\phi : B \to B$ (the minimum is attained when $\phi(b)$ equals the average $\bar{a}_t^{\mathbf{b}}(b)$ of the b-bin; cf. the correction of forecasts "in retrospect" in the Introduction).

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

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

where for each c in C we take $\bar{a}_0(c)$ to be an arbitrary element of C. Comparing this 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[9] (an average that will be determined only at time t, i.e., offline), by the

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

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*

$$\mathcal{R}_t \leq \widetilde{\mathcal{R}}_t \leq \mathcal{R}_t + \gamma^2 \frac{N_t}{t}\left(\ln\left(\frac{t}{N_t}\right) + 1\right), \tag{7}$$

where $N_t := |\{c_s : 1 \leq s \leq t\}|$ is the number of distinct elements in the sequence $\mathbf{c}_t = (c_1, ..., c_t)$ (i.e., the number of distinct forecasts used).

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[10] 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

$$0 \leq \widetilde{\mathcal{R}}_t - \mathcal{R}_t \leq |D| \frac{\ln t + 1}{t}. \tag{8}$$

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*

$$\sum_{i=1}^{n} \|x_i - \bar{x}_n\|^2 = \sum_{i=1}^{n} \left(1 - \frac{1}{i}\right) \|x_i - \bar{x}_{i-1}\|^2, \tag{9}$$

where $\bar{x}_m := (1/m)\sum_{i=1}^{m} x_i$ denotes the average of[11] *$x_1, ..., x_m$.*

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

$$s_n = s_{n-1} + \left(1 - \frac{1}{n}\right) \|x_n - \bar{x}_{n-1}\|^2. \tag{10}$$

We provide a short proof:[12] 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

[10] 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 $(a_{2i-1} - \bar{a}_{2i-2}(c_{2i-1}))^2 \geq 0$).

[11] The sum on the right-hand side of (9) effectively starts from $i = 2$, and so it does not matter how $\bar{x}_0$ is defined.

[12] An alternative proof of (10) uses $\mathbb{V}ar(X) = \mathbb{E}[\mathbb{V}ar(X|Y)] + \mathbb{V}ar(\mathbb{E}[X|Y])$, where $X = x_i$ with probability $1/n$ and Y is the indicator that $i = n$. Formula (10) is known as a "variance update" formula; see, e.g., Welford (1962).

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 (10) 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 (take $\bar{x}_0$ to be an arbitrary element of the convex hull of the x_i). 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

$$0 \le \widetilde{v}_n - v_n \le \frac{1}{n} \sum_{i=1}^n \frac{1}{i} \xi^2 \le \xi^2 \frac{\ln n + 1}{n}, \tag{11}$$

where $\xi := \max_{1 \le i,j \le n} \|x_i - x_j\|$; moreover, the[13] $O(\log n/n)$ bound is tight (take each x_i to be at a distance of at least some $\delta > 0$ from $\bar{x}_{i-1}$), and thus so is the $O(\log t/t)$ bound in Proposition 1 and (8).

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 (11) 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 (7). □

4 A Simple Way to Calibeat

We provide a simple calibeating procedure. The set B is taken for now to be finite (the restriction on the number of possible forecasts, i.e., on the number of bins, is needed

[13] We use standard asymptotic notation as $n \to \infty$: $f(n) = O(g(n))$, $f(n) = o(g(n))$, and $f(n) \sim g(n)$ stand for, respectively, $\overline{\lim}_{n\to\infty} f(n)/g(n) < \infty$, $\lim_{n\to\infty} f(n)/g(n) = 0$, and $\lim_{n\to\infty} f(n)/g(n) = 1$.

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, we have $\mathcal{R}_t = 0$ for all t). This finiteness assumption may be relaxed; see Remark (d) below.

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

$$c_t = \bar{a}_{t-1}^{\mathbf{b}}(b_t) \tag{12}$$

for every time $t \geq 1$ (if t is the first time that b_t is used, take c_t to be an arbitrary element of C). Then ζ is B-calibeating; specifically,

$$0 \leq \mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}} \leq \gamma^2 |B| \frac{\ln t + 1}{t} \tag{13}$$

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

Proof. Our choice of $c_t = \bar{a}_{t-1}^{\mathbf{b}}(b_t)$ makes $\mathcal{B}_t^{\mathbf{c}} = \widetilde{\mathcal{R}}_t^{\mathbf{b}}$ for any $\mathbf{a}_t$ and $\mathbf{b}_t$; use Proposition 1 (see (8)).[14] □

The calibeating forecast c_t is thus the average of the actions in those periods $1 \leq s \leq t-1$ in which the forecast b_s was equal to the current forecast b_t. When B is a subset of C we get by (13) that $\mathcal{B}_t^{\mathbf{c}} \leq \mathcal{B}_t^{\mathbf{b}} - \mathcal{K}_t^{\mathbf{b}} + o(1)$; 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 $t \geq 1$ and every $\mathbf{b}_t \in B^t$, we have

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

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 $N_t^{\mathbf{b}} := |\{b_s : s \leq t\}|$ is the number of distinct forecasts used by $\mathbf{b}$ up to time t (use Proposition 1).

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

(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 **b**-forecasts be generated by a forecasting procedure σ, and let σ' replace each b_t by the corresponding $\bar{a}^{\mathbf{b}}_{t-1}(b_t)$ (see Appendix A.1 for some technical details); then σ' yields lower Brier scores than σ in the long run: $\mathcal{B}^{\mathbf{c}}_t \leq \mathcal{B}^{\mathbf{b}}_t - \mathcal{K}^{\mathbf{b}}_t + o(1)$.

(f) The existence of a calibeating procedure may be proved by a minimax argument, which extends the 1995 proof of Hart of calibration (see Section 4 of Foster and Vohra 1998, and Hart 2021); we do so in Appendix A.2 of Foster and Hart (2022). Such existence proofs do not however provide explicit calibeating procedures, for sure not the very simple one of Theorem 3.

Appendix A.1 provides some further details. In particular, in Appendix A.1.1 we show that one cannot guarantee a calibeating error of an order of magnitude lower than $\log t/t$, 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 (12).

5 Self-calibeating = Calibrating

The construction of Section 4 may be leveraged to obtain calibration. Indeed, when $\mathbf{b} = \mathbf{c}$ we have $\mathcal{B}^{\mathbf{c}}_t - \mathcal{R}^{\mathbf{b}}_t = \mathcal{B}^{\mathbf{c}}_t - \mathcal{R}^{\mathbf{c}}_t = \mathcal{K}^{\mathbf{c}}_t$, and so "self-calibeating," i.e., **c** calibeating **c**, is equivalent to calibration, i.e., $\mathcal{K}^{\mathbf{c}}_t \to 0$. To achieve this by the construct of Theorem 3 we would need to choose c_t so that $c_t = \bar{a}^{\mathbf{c}}_{t-1}(c_t)$. However, this requires a fixed point of the function $\bar{a}^{\mathbf{c}}_{t-1}(\cdot)$, which of course need not exist in general. We circumvent this by using a "stochastic expected fixed point" result, 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).[15]

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 that is δ-calibrated; specifically,*[16]

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

for all $t \geq 1$ and all sequences $\mathbf{a}_t \in A^t$. Moreover, σ may be taken to be δ-almost deterministic (i.e., all randomizations are δ-local).

[15] 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 the proof more transparent.

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

Proof. For every t and history $h_{t-1} = (\mathbf{a}_{t-1}, \mathbf{c}_{t-1})$, the outgoing Theorem 10 (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

$$\mathbb{E}_{t-1}\left[\|a_t - c_t\|^2 - \|a_t - \bar{a}_{t-1}(c_t)\|^2\right] \leq \delta^2 \tag{14}$$

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 $t = 1, 2, \ldots$ 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 10. □

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 defined only on the finite set $\{c_s : 1 \leq s \leq t\}$ and is far from being continuous, 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).

Remarks. *(a)* From inequality (14) 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., $D_t \subseteq D_{t+1}$ for all t) 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

$$\bar{a}_t^{\mathbf{u}}(u) \equiv \bar{a}_t^{\mathbf{b}^1,\mathbf{b}^2}(b^1,b^2) := \frac{\sum_{1\le s\le t: u_s = x} a_s}{|\{1 \le s \le t : u_s = u\}|}$$

for every $u \in U$, 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; see Appendix A.4 for a formal proof (informally, 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 $\sum'(x_j - \bar{x}')^2 \le \sum'(x_j - \bar{x})^2$ [this holds for any y in place of $\bar{x}$], and similarly for $\sum''$, which added together yields $\sum'(x_j - \bar{x}')^2 + \sum''(x_j - \bar{x}'')^2 \le \sum(x_j - \bar{x})^2$). Therefore

$$\mathcal{R}_t^{\mathbf{b}^1,\mathbf{b}^2} \le \mathcal{R}_t^{\mathbf{b}^1} \quad \text{and} \quad \mathcal{R}_t^{\mathbf{b}^1,\mathbf{b}^2} \le \mathcal{R}_t^{\mathbf{b}^2}. \tag{15}$$

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 that is (δ, B)-calibeating and δ-calibrated; specifically,*

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

and thus, by (15),

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

for all $t \ge 1$ 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.

Thus, if we ignore 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), it 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 10 (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, by using it as the distribution of c_t given $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{b}_t)$ (which makes it a $\mathbf{b}$-based

procedure), 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 $a_t \in A$, where $\mathbb{E}_{t-1}$ denotes the expectation conditional on $(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, \mathbf{b}_t)$. 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 above proof amounts to using a stochastic forecast-hedging calibration procedure separately for 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, which may well be quickly overwhelmed by the accumulating errors of calibeating, as well as those due to rounding up to a finite grid.

(d) A proof that is directly based on the minimax theorem is provided in Appendix A.2 of Foster and Hart (2022).

Calibration, and thus calibeating by a calibrated forecast, requires the procedure to be stochastic. However, if we replace calibration with *continuous calibration*, a weakening defined in Foster and Hart (2021)—useful, in particular, for game dynamics that yield Nash equilibria—we get a *deterministic* procedure instead.

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

We relegate the details to Appendix A.6.

7 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 7 (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 (if t is the first period in which $(b_t^1, ..., b_t^N)$ is used, take $c_t \in C$ to be arbitrary).*

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

(iii) *There exists a deterministic $(\mathbf{b}^1, ..., \mathbf{b}^N)$-based C-forecasting procedure ζ 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 a multi-calibeating procedure that is more complex but yields a smaller error term (see also Appendix A.8.1 in Foster and Hart 2022).

(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 (ii) 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 omitted proofs and several sharpenings of our results.

A.1 A Simple Way to Calibeat

First, we provide the technical details for Remark (e) in Section 4. Assume that the **b**-forecasts are generated by a forecasting procedure σ; then the procedure σ', 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 $\Gamma_{t-1} := \{\mathbf{b}_{t-1} : \bar{a}^{\mathbf{b}}_{s-1}(b_s) = c_s$ for every $1 \leq s \leq t-1\}$ and $\Gamma_t := \{\mathbf{b}_t : \bar{a}^{\mathbf{b}}_{s-1}(b_s) = c_s$ for every $1 \leq s \leq t\}$ be the sets of $\mathbf{b}_{t-1}$ and $\mathbf{b}_t$ 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

$$\sigma''(\mathbf{a}_{t-1}, \mathbf{c}_{t-1})(c_t) := \frac{\sum_{\mathbf{b}_t \in \Gamma_t} \prod_{s=1}^{t} \sigma(\mathbf{a}_{s-1}, \mathbf{b}_{s-1})(b_s)}{\sum_{\mathbf{b}_{t-1} \in \Gamma_{t-1}} \prod_{s=1}^{t-1} \sigma(\mathbf{a}_{s-1}, \mathbf{b}_{s-1})(b_s)}. \tag{17}$$

Second, we address the question of whether one can do better than by choosing $c_t = \bar{a}^{\mathbf{b}}_{t-1}(b_t)$ at each time t in Theorem 3. Since $\widetilde{\mathcal{R}}^{\mathbf{b}}_t - \mathcal{R}^{\mathbf{b}}_t = O(\log t/t) \to 0$, consider the game where our **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

$$\mathcal{B}^{\mathbf{c}}_t - \widetilde{\mathcal{R}}^{\mathbf{b}}_t = \frac{1}{t} \sum_{s=1}^{t} \left[||a_s - c_s||^2 - \left\| a_s - \bar{a}^{\mathbf{b}}_{s-1}(b_s) \right\|^2 \right]. \tag{18}$$

Suppose that we are in a period $r \leq t$, and so the terms $s < r$ of the sum (18) are all given. To *guarantee*—no matter 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* now choose $c_r = \bar{a}^{\mathbf{b}}_{r-1}(b_r)$. This follows since for every $\bar{a} \in \text{conv}A$ and $c \neq \bar{a}$ we have[17]

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

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

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

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

A.1.1 The Calibeating Error

What is the minimal calibeating error that can be guaranteed? First, we show (Proposition 8) 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 (Proposition 9): 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.

Proposition 8 shows that one cannot guarantee a calibeating error of an order of magnitude lower than $\log t/t$, even in the simplest one-dimensional case; see Remark (a) below for the extension to the multidimensional case.

Proposition 8 *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*[18]

$$\sup_{\mathbf{a}_t \in A^t} \mathbb{E}\left[\mathcal{B}^{\mathbf{c}}_t - \mathcal{R}^{\mathbf{b}}_t\right] \geq \left(\frac{1}{4} - o(1)\right) \frac{\ln t}{t} \tag{19}$$

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 $\mathcal{B}^{\mathbf{c}}_t - \mathcal{R}^{\mathbf{b}}_t$

[17] One way to see this is as follows. Let $\bar{a} = \sum_i \lambda_i a_i$ be a convex combination of elements a_i in A; 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 i we must have $\|a_i - c\|^2 \geq \|a_i - \bar{a}\|^2 + \|c - \bar{a}\|^2$.

[18] For a constant sequence $\mathbf{b}$, a $\mathbf{b}$-based forecasting procedure is simply a forecasting procedure. The expectation in (19) is over the stochastic choices of ζ (and it applies only to $\mathcal{B}^{\mathbf{c}}_t$, since $\mathcal{R}^{\mathbf{b}}_t$ is determined by $\mathbf{a}_t$ alone when $\mathbf{b}_t$ is a constant sequence).

(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

$$\inf_{\zeta} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \geq \left(\frac{1}{4} - o(1)\right) \frac{\ln t}{t}, \tag{20}$$

where the infimum is taken over all forecasting procedures ζ (and the expectation is over the randomizations of both actions and forecasts). 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 (19).

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):

$$\mathbb{E}\left[\bar{a}_t\right] = \frac{1}{2} \quad \text{and} \quad \mathbb{V}ar\left[\bar{a}_t\right] = \mathbb{V}ar\left[\bar{a}_t\right] = \frac{t + 2\alpha}{4(2\alpha + 1)t}; \tag{21}$$

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 (21),

$$\mathbb{E}\left[\hat{\theta}_t\right] = \frac{1}{2} \quad \text{and} \quad \mathbb{V}ar\left[\hat{\theta}_t\right] = \frac{t}{4(2\alpha + 1)(t + 2\alpha)}. \tag{22}$$

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 (21), gives

$$\begin{aligned} \mathbb{E}\left[\mathcal{R}_t^{\mathbf{b}}\right] &= \mathbb{E}\left[\bar{a}_t(1 - \bar{a}_t)\right] = \mathbb{E}\left[\bar{a}_t\right] - \mathbb{E}^2\left[\bar{a}_t\right] - \mathbb{V}ar\left[\bar{a}_t\right] \\ &= \frac{1}{2} - \frac{1}{4} - \frac{t + 2\alpha}{4(2\alpha + 1)t} = \lambda - \frac{\lambda}{t}, \end{aligned} \tag{23}$$

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[19] X,

[19] 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 nonposi-

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 (22) 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 (23) 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[20] $1/4$ by taking large enough α we get (20), which completes the proof. □

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

$$\sup_{\mathbf{a}_t \in A^t} \mathbb{E}\left[\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^{\mathbf{b}}\right] \geq \left(\frac{m}{4} - o(1)\right)\frac{\ln t}{t} \tag{24}$$

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 8 to the subsequence where $b_t = b$ for each $b \in B$ separately and then summing up yields

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

We now show that one can lower 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 Proposition 8.

Assuming that C is a full-dimensional set in[21] $\mathbb{R}^m$, 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

tively 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 $\mathbb{C}ov\left[X,Y\right] \leq 0$.

[20] 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 $\log t/t$.

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

relation of r to the diameter γ of C is, by Jung's (1901) theorem,

$$r^2 \le \gamma^2 \frac{m}{2(m+1)} \tag{25}$$

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

Proposition 9 *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 $t \ge 1$. Then ζ is B-calibeating, and

$$\mathcal{B}_t^{\mathbf{c}'} - \mathcal{R}_t^{\mathbf{b}} \le r^2 |B| \frac{\ln t + 1}{t} \tag{26}$$

for all $t \ge 1$ 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 (13), the upper bound of (26) on the calibeating error has r^2 instead of γ^2, which, by (25), 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., (12).

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

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

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

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

(because $|B_t| \le |B|$ and $n_t^{\mathbf{b}}(b) \le t$), which is (26). □

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

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

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 (24).

A.2 Complexity of Procedures: Minimax (MM) and Fixed Point (FP) Procedures

The basic calibeating procedure of Theorem 3 is very simple, as it requires just the computation of averages; the same holds for the multi-calibeating procedure of Theorem 7 (i). The other procedures that we provide are more complex, and require solving at each step a certain multidimensional problem. These problems turn out to be of two distinct kinds: for stochastic procedures, they are finite minimax (or linear programming) problems, and for deterministic and almost-deterministic procedures, they are continuous fixed point problems. (The existence of the corresponding solutions is proven by the von Neumann 1928 minimax theorem and the Brouwer 1912 fixed point theorem, respectively; see Appendix A.3 below.) Following Foster and Hart (2021, Section III.D), we refer to these as being of *type MM* (minmax) and *type FP* (fixed point), respectively.

This distinction, which is significant in the multidimensional case (i.e., for $m > 1$) and is of the polynomial vs. nonpolynomial variety, is not just a matter of proof technique; see Foster and Hart (2021), Sections III.D, VI, and VII (with a summary in Table I there). Theorem 10 in Appendix A.3, which provides the "outgoing" tools that we use, makes the distinction clear: part (S) gives procedures of type MM, and parts (D) and (AD) give procedures of type FP.

Specifically, the results that yield procedures of type MM are: calibration (Theorem 4), calibeating with calibration (Theorem 5), and multi-calibeating with calibration (Theorem 7 (ii))—all of them without the "moreover" almost-deterministic statement. The results that yield FP procedures are all the above "moreover" statements, calibeating with continuous calibration, and multi-calibeating with continuous calibration (Theorems 6, 12, and 7 (iii)).

A.3 "Outgoing" Results

We provide the results of the "outgoing" theorems of Foster and Hart (2021), restating them in a convenient manner for our use. The seemingly slightly weaker formulations here are still equivalent to Brouwer's fixed point theorem and von Neumann's minimax theorem, respectively; see Appendix A.4 in the long version of the paper, Foster and Hart (2022). The FP vs. MM distinction is discussed in Appendix A.2 above. 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 10 *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*

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

(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*

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

(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

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

with $z = g(y)$, to get 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

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

and thus $\|x - y\|^2 - \|x - g(y)\|^2 \le 0$, 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

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

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

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

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

(AD) This is the same proof as for (S), except that it uses 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 fixed point theorem) then condition (27), which says that $\lambda \le 0$ 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 $\lambda = \|g(y) - y\|^2 \le 0$, and thus $g(y) = y$). Condition (28) extends this by requiring that $\lambda \le 0$ hold approximately on average, i.e., $\mathbb{E}[\lambda] \le \delta^2$, 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; this example also shows that one cannot strengthen $\mathbb{E}[\lambda] \leq \delta^2$ to $\mathbb{E}[|\lambda|] \leq \delta^2$ —i.e., "$\lambda = 0$" instead of "$\lambda \leq 0$"—because for $x = 0$ we have $|\lambda| = |y^2 - g(y)^2| \geq 1/4$ for all $y \in C$).

Appendix A.4 of Foster and Hart (2022) contains additional relevant material.

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 $\mathcal{R}_t^{\mathbf{b}^1,...,\mathbf{b}^N} \leq \mathcal{R}_t^{\mathbf{b}^n}$ for each $n = 1, ..., N$ (Section 7) and also $\mathcal{R}_t^{\mathbf{b},\Pi} \leq \mathcal{R}_t^{\mathbf{b}}$ and $\mathcal{R}_t^{\mathbf{b},\Pi} \leq \mathcal{R}_t^{\Pi}$ (Appendix A.6).

We consider general fractional binnings. Let I be a finite or countably infinite collection of bins, and consider a sequence $(z_s)_{s\geq 1}$ (namely, $z_s = a_s - c_s$) such that at time s the fraction $\lambda_s(i) \geq 0$ 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 $t \geq 1$ (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 i (when $\sum_{s\leq t} \lambda_s(i) > 0$).

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 $\sum_{s\leq t} \sum_{i\in I} \lambda_s(i)/t = 1$); thus, $\mathbb{P}[(Z,U) = (z,i)] = (1/t) \sum_{s\leq t: z_s = z} \lambda_s(i)$, which is the average, over all periods $s = 1, ..., t$, of the probability that the value z goes into bin i. We then have

$$\begin{aligned}
\mathbb{P}[U = i] &= \sum_{s=1}^{t} \frac{\lambda_s(i)}{t}, \\
\mathbb{E}[Z|U = i] &= \frac{1}{\mathbb{P}[U = i]} \sum_{s=1}^{t} \left(\frac{\lambda_s(i)}{t} \right) z_s \;=\; \bar{z}(i), \\
\mathbb{V}ar[Z|U = i] &= \frac{1}{\mathbb{P}[U = i]} \sum_{s=1}^{t} \left(\frac{\lambda_s(i)}{t} \right) (z_s - \bar{z}(i))^2, \text{and} \\
\mathbb{E}[\mathbb{V}ar[Z|U]] &= \sum_{i\in I} \mathbb{P}[U = i]\, \mathbb{V}ar[Z|U = i] \;=\; \mathcal{R}. \qquad (31)
\end{aligned}$$

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

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

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

$$\mathbb{E}\left[\mathbb{V}ar\left[Z|\mathcal{F}_1\right]\right] \geq \mathbb{E}\left[\mathbb{V}ar\left[Z|\mathcal{F}_2\right]\right], \tag{32}$$

which yields the result by (31).

Applying the classic inequality $\mathbb{V}ar\left[X\right] = \mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|^2\right] \leq \mathbb{E}\left[\|X - x\|^2\right]$ 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.)

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

because $\mathbb{E}\left[Z|\mathcal{F}_1\right]$ 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 $\mathbb{E}\left[\|Z - \mathbb{E}\left[Z|\mathcal{F}_1\right]\|^2 |\mathcal{F}_1\right]$ (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 (32), and thus completes the proof. □

Applying Proposition 11 with ϕ being a projection, such as $\phi(b^1, ..., b^N) = b^n$, yields the needed inequalities.

A.5 General Brier Score Decomposition

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

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

(we use this in the proof of Theorems 6 and 12). Indeed, 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 Calibeating by a Deterministic Continuously Calibrated Forecast

In this appendix we prove Theorem 6 in Section 6: one can guarantee calibeating by a *deterministic* procedure that is *continuously calibrated*, a useful weakening of calibration (see Foster and Hart 2021).

We start by recalling 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

$$\lim_{t\to\infty} \left(\sup_{\mathbf{a}_t} \mathcal{K}_t^{\Pi} \right) = 0 \tag{33}$$

for every continuous binning Π, where the Π-*calibration* score $\mathcal{K}_t^{\Pi}$ is[22]

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

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 (33) for one specific continuous binning Π_0; i.e., σ is continuously calibrated if and only if (33) holds for $\Pi = \Pi_0$.

Let B be an arbitrary finite set[23] 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 $\mathbf{1}_x(y) = 0$

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

[23] 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).

for $y \neq x$). Consider bin (b,i) at time t; its total weight, average, and variance are, 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 now state a more detailed version of Theorem 6 on calibeating by a *deterministic* continuously calibrated procedure.

Theorem 12 *Let B be a finite set. Then there exists a* deterministic $\mathbf{b}$*-based forecasting procedure ζ that is B-calibeating and is continuously calibrated; specifically,*

$$\mathcal{B}_t^{\mathbf{c}} \le \mathcal{R}_t^{\mathbf{b},\Pi_0} + o(1), \tag{34}$$

where Π_0 is the continuous binning given by Proposition 3 in Foster and Hart (2021), which 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 13 *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 **a**, **b**, *and* **c**.

The proof is an adaptation of the proof of Proposition 1 to fractional binnings. We start by generalizing 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 $\Lambda_n := \sum_{i=1}^n \lambda_i$. Let $\bar{x}_n := \sum_{i=1}^n (\lambda_i/\Lambda_n) x_i$ denote the weighted average of $x_1, ..., x_n$ (when $\Lambda_n = 0$, and thus $\lambda_i = 0$ for all $i = 1, ..., n$, put $\lambda_i/\Lambda_n = 0/0 = 0$).

Proposition 14 *For every $n \geq 1$ we have*[24]

$$\sum_{i=1}^n \lambda_i \left\|x_i - \bar{x}_n\right\|^2 = \sum_{i=1}^n \lambda_i \left(1 - \frac{\lambda_i}{\Lambda_i}\right) \left\|x_i - \bar{x}_{i-1}\right\|^2. \tag{35}$$

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

$$s_n = s_{n-1} + \lambda_n \left(1 - \frac{\lambda_n}{\Lambda_n}\right) \left\|x_n - \bar{x}_{n-1}\right\|^2. \tag{36}$$

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 (36) 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 $x_1, ..., x_n$ (again, take $\bar{x}_0$ to be an arbitrary element of the convex hull of the x_i). Proposition 14 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),

$$0 \leq \widetilde{v}_n - v_n \leq \xi^2 \frac{1}{\Lambda_n} \sum_{i=1}^n \frac{\lambda_i^2}{\Lambda_i} \leq \xi^2 \frac{\ln \Lambda_n + 2}{\Lambda_n}, \tag{37}$$

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

We now prove Proposition 13, which shows that the online refinement score $\widetilde{\mathcal{R}}_t^{\mathbf{b},\Pi}$ is close to the (offline) refinement score.

[24] The sum on the right-hand side of (35) effectively starts from $i = 2$, and so, as in (9), it does not matter how $\bar{x}_0$ is defined.

Proof of Proposition 13. 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

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

for each $(b,i)\in B\times I$. Proposition 14, specifically, (37), yields

$$0 \le \mu_t(b,i) \le 4\gamma^2(\ln n_t(b,i) + 2) \le 4\gamma^2(\ln t + 2) \tag{38}$$

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

$$0 \le \frac{1}{t}\sum_{b\in B}\sum_{i\in J}\mu_t(b,i) \le 4\gamma^2|B|\;|J|\;\frac{\ln t + 2}{t}. \tag{39}$$

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 $\sum_{i\in I\setminus J} w_i(c)\le\varepsilon$ for all $c\in C$; such a finite J exists by Dini's theorem (see (9) in Foster and Hart 2021). For $i\in I\setminus J$ we get

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

Adding this to (39) yields

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

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 $\mathbf{c}$. □

Finally, we prove Theorem 12 (and thus Theorem 6).

Proof of Theorem 12. 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 result, specifically, Theorem 10 (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

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

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 (34) by Proposition 13. 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). □

A.7 Multi-calibeating: Improved Error Term

The error term of the multi-calibeating procedure of Theorem 7 has a constant of $\prod_{n=1}^{N} |B^n|$, which increases exponentially with the number of given procedures N. We provide here an alternative construction, based on "online linear regression" methods (see Azoury and Warmuth 2001), whose constant, $N + \max_n |B_n|$, increases linearly with N.

We will use 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 $\bar{a}_{t-1}^{\mathbf{b}^n}(b_t^n)$. Assume that $C \subseteq [-\gamma_0, \gamma_0]^m$ (see remark (a) 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 $x_{i,t}^n := \bar{a}_{i,t-1}^n(b_t^n)$ (this is the average of the ith coordinates of a_s over all periods $s \leq t$ in which $b_s^n = b_t^n$), and put $x_{i,t} := (x_{i,t}^n)_{n=1,...,N} \in \mathbb{R}^N$.

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

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

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

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

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

$$\sum_{i=1}^{m} \mathcal{F}_{i,t}^* \leq \sum_{i=1}^{m} \mathcal{F}_{i,t}(e^n) \leq \widetilde{\mathcal{R}}_t^n + \frac{m\alpha}{t}. \tag{40}$$

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

$$\sum_{s=1}^{t}(a_{i,s} - \theta_{i,s} \cdot x_{i,s})^2 \leq t\mathcal{F}_{i,t}^* + \gamma_0 N \ln\left(\frac{\gamma_0}{\alpha}t + 1\right) \tag{41}$$

is guaranteed for any sequence[25] $(a_{i,s})_{s \geq 1}$.

[25] This is Theorem 5.6 of Azoury and Warmuth (2001); in the notation there, $X = \max_{n,s} |x_{i,s}^n| \leq \gamma_0$ and $Y = \max_s |a_{i,s}| \leq \gamma_0$. Note that there is a misprinted sign in the first line of formula (5.17) there.

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 $c_s := \text{proj}_C(\hat{c}_s)$ 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

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

Averaging over $s = 1, ..., t$ gives

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

by (41). Recalling (40) and Proposition 1 yields

$$\mathcal{B}_t^{\mathbf{c}} - \mathcal{R}_t^n \leq \frac{m\gamma_0 N}{t} \ln\left(\frac{\gamma_0}{\alpha} t + 1\right) + \frac{m\alpha}{t} + \gamma^2 |B^n| \frac{\ln t + 1}{t} = O\left(\left(N + \max_n |B_n|\right) \frac{\log t}{t}\right). \tag{42}$$

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 $\gamma/2 \leq r \leq \gamma\sqrt{m/(2m+2)}$ 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) 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 $\mathcal{F}_{i,t}(\theta)$ 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.

(c) Since we use the inequalities $\mathcal{F}_{i,t}^* \leq \mathcal{F}_{i,t}(\theta)$ only for θ equal to the unit vectors e^n in $\mathbb{R}^N$, it suffices to minimize $\mathcal{F}_{i,t}(\theta)$ over the convex hull of these vectors, that is, over the unit simplex $\Delta(N)$ of $\mathbb{R}^N$, as in Foster (1991) (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 multi-calibeating is equivalent to being, in terms of the Brier scores, "as strong as" each one of the N sequences $(\bar{a}_{t-1}^n(b_t^n))_{t \geq 1}$ (for $n = 1, ..., N$).

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

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

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

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

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

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

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

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

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

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

Foster, Dean P. and Sergiu Hart (2021), "Forecast Hedging and Calibration," *Journal of Political Economy* 129, 3447–3490.

Foster, Dean P. and Sergiu Hart (2022), " 'Calibeating': Beating Forecasters at Their Own Game," `http://arxiv.org/abs/2209.04892v2` [long version of the present paper]

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

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

Hart, Sergiu (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, Norman L., Adrienne W. Kemp, and Samuel Kotz (2005), *Univariate Discrete Distributions*, 3rd edition, Wiley.

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

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

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

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

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

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

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

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

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

Shmaya, Eran (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, Vladimir (2001), "Competitive On-Line Statistics," *International Statistical Review* 69, 213–248.

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