Stronger Calibration Lower Bounds via Sidestepping

Mingda Qiao
Stanford University
Stanford, CA, USA
mqiao@stanford.edu

Gregory Valiant
Stanford University
Stanford, CA, USA
valiant@stanford.edu

ABSTRACT

We consider an online binary prediction setting where a forecaster observes a sequence of T bits one by one. Before each bit is revealed, the forecaster predicts the probability that the bit is 1. The forecaster is called well-calibrated if for each $p \in [0, 1]$, among the n_p bits for which the forecaster predicts probability p, the actual number of ones, m_p, is indeed equal to $p \cdot n_p$. The *calibration error*, defined as $\sum_p |m_p - pn_p|$, quantifies the extent to which the forecaster deviates from being well-calibrated. It has long been known that an $O(T^{2/3})$ calibration error is achievable even when the bits are chosen adversarially, and possibly based on the previous predictions. However, little is known on the lower bound side, except an $\Omega(\sqrt{T})$ bound that follows from the trivial example of independent fair coin flips.

In this paper, we prove an $\Omega(T^{0.528})$ bound on the calibration error, which is the first super-$\sqrt{T}$ lower bound for this setting to the best of our knowledge. The technical contributions of our work include two lower bound techniques, *early stopping* and *sidestepping*, which circumvent the obstacles that have previously hindered strong calibration lower bounds. We also propose an abstraction of the prediction setting, termed the Sign-Preservation game, which may be of independent interest. This game has a much smaller state space than the full prediction setting and allows simpler analyses. The $\Omega(T^{0.528})$ lower bound follows from a general reduction theorem that translates lower bounds on the game value of Sign-Preservation into lower bounds on the calibration error.

CCS CONCEPTS

• **Theory of computation → Online learning algorithms**.

KEYWORDS

online prediction, calibration

ACM Reference Format:
Mingda Qiao and Gregory Valiant. 2021. Stronger Calibration Lower Bounds via Sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21), June 21–25, 2021, Virtual, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3406325.3451050

STOC '21, June 21–25, 2021, Virtual, Italy

ACM ISBN 978-1-4503-8053-9/21/06…$15.00
https://doi.org/10.1145/3406325.3451050

1 INTRODUCTION

We study the following online binary prediction problem. A forecaster predicts a binary sequence of length T that is observed one bit at a time. Before seeing each bit, the forecaster makes a prediction about the probability that this bit is a "1". For simplicity, we require all the predictions to fall in a finite set $P \subset [0, 1]$ specified by the forecaster at the beginning. At the end of the T time steps, the *calibration error*[1] incurred by the forecaster is defined as

$$\mathrm{calerr}(T) \coloneqq \sum_{p \in P} |m_p(T) - n_p(T) \cdot p|, \tag{1}$$

where $n_p(T)$ denotes the number of times probability p is predicted up to time T, and $m_p(T)$ is the number of ones observed among those $n_p(T)$ time steps. Thus, the calibration error quantifies the extent to which the forecaster's predictions are well-calibrated in the sense that for every possible prediction value p, the frequency of 1 among the time steps at which p is predicted is indeed close to p.

The notion of calibration is incomparable to other usual performance metrics such as prediction accuracy. Particularly when predicting potentially noisy binary outcomes, it is difficult to establish good benchmarks for prediction accuracy, as it is generally impossible to distinguish between noise in the observations, versus a failure of the predictor. By contrast, calibration is a natural desiderata that applies whether or not the observations have intrinsic noise. Being well-calibrated can be viewed as a minimum requirement on the forecaster for its predictions to be interpreted as meaningful probabilities: Suppose that among all the days on which a weather forecast predicted a 50% chance of rain, it rained on only 10% of the days. The predictions of such weather forecasts clearly lack credibility.

There has been a recent surge of interest in calibration, both from the machine learning community (e.g., [8, 13, 14, 21]), and from the perspective of algorithmic fairness (e.g., [10–12, 16, 17, 19]). In the machine learning setting, this study is motivated in part by the fact that trained neural networks are often spectacularly poorly calibrated and overconfident in their predictions. The connection between calibration and fairness is especially natural: as proposed in [12], a fair classifier should be calibrated on every protected demographic group. Indeed, predictions that are not well-calibrated for some demographic groups would seem to conflict with the most intuitive notion of what it means to treat all groups fairly.

The calibration aspect of online predictions was first studied by Foster and Vohra [6]. They gave a randomized forecaster that

[1] More generally, the ℓ_q calibration error is defined as $\left(\sum_{p \in P} \frac{n_p(T)}{T} \left| \frac{m_p(T)}{n_p(T)} - p \right|^q \right)^{1/q}$ in the literature. Up to a factor of T, the definition in (1) coincides with the ℓ_1 calibration error, which is also called the expected calibration error (ECE).

achieves an $O(T^{2/3})$ calibration error in expectation, even if the T bits are chosen by an adaptive adversary that chooses the t-th bit based on the bits and predictions in the previous $t-1$ steps. This $O(T^{2/3})$ upper bound has a simple non-constructive proof due to Sergiu Hart based on the minimax theorem [6, Section 4]: For each fixed finite set P, each deterministic adaptive strategy of the forecaster (resp. the adversary) can be viewed as a function that maps $\bigcup_{t=0}^{T-1}(\{0,1\}^t \times P^t)$ to P (resp. $\{0,1\}$), so there are only finitely many such strategies. Thus, by the minimax theorem, it suffices to prove the following claim: Against any given adversary (which might be randomized and adaptive), there is a forecaster that achieves an $O(T^{2/3})$ calibration error in expectation. This claim, in turn, can be proved using the following "rounding strategy": (1) choose $P = \left\{0, \frac{1}{T^{1/3}}, \frac{2}{T^{1/3}}, \ldots, 1\right\}$; (2) at each time step t, compute the probability of the event $b(t) = 1$ conditioning on the previous $t-1$ steps (using the knowledge of the given adversary), and predict the value in P that is closest to this conditional probability. We refer the readers to [9] for further details of this proof.

On the other hand, less is understood on the lower bound side. The only known lower bound on the calibration error is $\Omega(\sqrt{T})$, which can be proved using a simple adversary that outputs T independent and uniformly distributed random bits. In this case, the optimal strategy is to predict probability $1/2$ at every step t. Then, the calibration error $\mathrm{calerr}(T)$ reduces to $|m_{1/2}(T) - T/2|$, where $m_{1/2}(T)$ follows the binomial distribution $B(T, 1/2)$, and this error is $\Omega(\sqrt{T})$ in expectation. Unfortunately, there is no known scheme of the adversary that outperforms the trivial one (that outputs independent coin flips) and gives a $\omega(\sqrt{T})$ bound on the calibration error, not to mention an $\Omega(T^{2/3})$ bound that matches the best known upper bound.

1.1 The Prediction Setting

The binary prediction setting is formally defined as a two-player multi-stage game between a forecaster and an adversary. The forecaster first specifies a finite set $P \subset [0,1]$ from which the predictions are selected. At each time step $t = 1, 2, \ldots, T$, the forecaster chooses $p(t) \in P$ and the adversary chooses $b(t) \in \{0,1\}$ simultaneously. Both choices may depend on the previous $t-1$ steps but not the other player's action at time t. For each $p \in P$, let $n_p(t) := \sum_{i=1}^{t} \mathbb{I}[p(i) = p]$ denote the number of times that p is predicted by the forecaster during the first t time steps, and let $m_p(t) := \sum_{i=1}^{t} \mathbb{I}[p(i) = p \wedge b(i) = 1]$ denote the number of time steps at which p is predicted and the bit chosen by the adversary is 1. Then, the cumulative calibration error up to time t is defined as

$$\mathrm{calerr}(t) := \sum_{p \in P} |m_p(t) - n_p(t) \cdot p|.$$

Define $\Delta_p(t) := m_p(t) - n_p(t) \cdot p$ as the total bias associated with prediction value p after the first t time steps. Moreover, let $\Delta_p^+(t) := \max(\Delta_p(t), 0)$ and $\Delta_p^-(t) := \max(-\Delta_p(t), 0)$ denote the positive and negative parts of $\Delta_p(t)$. Then, $\mathrm{calerr}(t)$ can be equivalently written as

$$\mathrm{calerr}(t) = \sum_{p \in P} |\Delta_p(t)| = \sum_{p \in P} \Delta_p^+(t) + \sum_{p \in P} \Delta_p^-(t).$$

For each quantity that is labeled by a time step (e.g., $\mathrm{calerr}(t)$ and $\Delta_p(t)$), we may omit the parameter t (and simply write, e.g., calerr and Δ_p) if it can be inferred from the context. In particular, we will drop the notation t when describing a scheme of the adversary, since the time t is naturally given by the time step at which that statement is executed.

It should be noted that the finiteness assumption on P is not too restrictive and is standard in the literature (e.g., [6]). This assumption can be justified by real-world prediction scenarios such as weather forecasts, where the probability of precipitation is typically rounded to 5% or 10% increments. Moreover, we can verify that rounding each prediction $p(t)$ to the nearest multiple of $1/T$ would increase $\mathrm{calerr}(T)$ by at most an additive constant. Thus, it is without loss of generality to take $P = \{0, 1/T, 2/T, \ldots, 1\}$.

1.2 Obstacles Against Strong Lower Bounds

Recall that an $O(T^{2/3})$ upper bound can be proved by analyzing a forecaster that predicts the nearest multiple of $1/T^{1/3}$ to the conditional probability that the next bit is 1. Suppose that the adversary divides the time horizon T into $k = T^{1/3}$ "epochs" of length T/k, and outputs T/k independent samples from the Bernoulli distribution $\mathrm{Ber}(i/k)$ in the i-th epoch. Then, the forecaster with the rounding strategy would keep predicting probability i/k in the i-th epoch, and the expected calibration error is given by

$$\begin{aligned}\mathbb{E}[\mathrm{calerr}(T)] &= \sum_{i=1}^{k} \mathop{\mathbb{E}}_{X \sim B(T/k, i/k)} [|X - (T/k) \cdot (i/k)|] \\ &= \Omega(k \cdot \sqrt{T/k}) = \Omega(T^{2/3}),\end{aligned}$$

where $B(\cdot,\cdot)$ denotes the binomial distribution. This indicates that the analysis of the $O(T^{2/3})$ upper bound is tight. Assuming that the forecaster is "truthful" in the sense that its prediction is always equal to (or very close to) the conditional probability of the next bit, the above example also suggests an $\Omega(T^{2/3})$ lower bound for all such truthful forecasters.

Unfortunately, this truthfulness assumption on the forecaster does not always hold; in various scenarios the forecaster, to minimize the calibration error, has an incentive to make untruthful predictions that are far away from the true probabilities. In the following, we describe several such scenarios, including *coarse-grained binning* and *cover-up*, that make the construction of lower bound schemes highly nontrivial.

The first example shows that, while the above construction proves the tightness of the upper bound analysis, there exists another simple forecaster that achieves a small error on it.

Example 1 (Coarse-grained binning). *Let us revisit the case that the binary sequence consists of T/k independent samples from each of $\mathrm{Ber}(1/k), \mathrm{Ber}(2/k), \ldots, \mathrm{Ber}(k/k)$ for $k = T^{1/3}$. Note that the sum of the T bits has an expectation of $T \cdot \frac{k+1}{2k}$ and an $O(T)$ variance. Therefore, if the forecaster predicts $\frac{k+1}{2k}$ at each of the T steps, the resulting calibration error is $O(\sqrt{T})$ in expectation.*

In Example 1, while we expect the forecaster to put the T bits into k "bins" associated with probabilities $1/k, 2/k, \ldots, k/k$ faithfully and incur an $\Omega(T^{2/3})$ error, the forecaster would actually merge all these bins into a larger, coarse-grained bin corresponding to

probability $\frac{k+1}{2k}$. More generally, as long as the T bits are independently drawn with fixed probabilities $p_1^*, p_2^*, \ldots, p_T^*$, the forecaster may as well predict the average $\frac{1}{T}\sum_{t=1}^{T} p_t^*$ at every single time step, resulting in $\mathbb{E}[\mathrm{calerr}(T)] = O(\sqrt{T})$.

Example 2 (Cover-up). *Suppose that the sequence consists of $T/3$ uniformly random bits followed by $T/3$ ones and then $T/3$ zeros. Moreover, suppose that the forecaster predicts $1/2$ in each of the first $T/3$ steps. Then, the calibration error after the first $T/3$ steps is $\Omega(\sqrt{T})$ in expectation. However, the forecaster can always "cover up" this error using the subsequent bits: If the first $T/3$ bits contain more zeros than ones, the forecaster may keep predicting $1/2$ (even though the bits are known to be ones) until $m_{1/2}(t) = n_{1/2}(t)/2$ at some point t. Similarly, the forecaster may cover up the error by predicting $1/2$ during the last $T/3$ time steps, if ones outnumber zeros among the first $T/3$ bits.*

In Example 2, the forecaster can always achieve a zero calibration error by untruthfully predicting $1/2$ for bits that are known to be deterministic. While the example might appear a bit contrived, this phenomenon that a forecaster can strategically decrease the cumulative calibration error by predicting untruthfully is not uncommon. Foster and Hart [4] refer to such behavior as "backcasting" (in contrast to forecasting), in the sense that the forecaster makes use of the future outcomes to disguise the mistakes it has made in the past.

The following example, termed "forecast hedging" in [4], indicates that the cover-up scenario is universal and makes it difficult to prove strong calibration lower bounds.

Example 3 (Forecast hedging). *Suppose that at time t, it holds that $\Delta_{p_1}(t) \le -1$ and $\Delta_{p_2}(t) \ge 1$ for some $p_1 < p_2$. We claim that the forecaster can decrease the calibration error in expectation after the next time step (i.e., ensure that $\mathbb{E}[\mathrm{calerr}(t+1)] \le \mathrm{calerr}(t)$) by predicting either p_1 or p_2, each with probability $1/2$.*

To see this, first suppose that the next bit is 0. Then, with probability $1/2$, Δ_{p_1} is decreased by p_1, and calerr, which contains a $|\Delta_{p_1}|$ term, will be increased by p_1; with the remaining probability $1/2$, Δ_{p_2} is decreased by p_2, and calerr also drops by p_2. In expectation, the cumulative calibration error drops by $\frac{p_2-p_1}{2} > 0$. A similar analysis works for the case that the next bit is 1. Thereby, the forecaster can cancel out part of the previous error by randomizing between the two predictions p_1 and p_2, without taking into account the actual probability of the next bit.

1.3 Our Results

The main result of this paper is the first super-$\sqrt{T}$ lower bound on the calibration error in the online binary prediction setting.

Theorem 1. *Let $\alpha = \frac{\log 8}{\log 255}$, $\beta = \frac{\log(9/2)}{\log 255}$, and $c = \frac{\beta+1}{\alpha+2\beta+1} > 0.528$. There exists a scheme of the adversary such that every forecaster incurs an $\Omega(T^c/\sqrt{\log T}) = \Omega(T^{0.528})$ calibration error in expectation.*

The proof of Theorem 1 builds on two simple yet powerful lower bound techniques tailored to calibration error, termed as *early stopping* and *sidestepping*, that manage to overcome the obstacles discussed in Section 1.2.

Early stopping. To prevent the forecaster from putting all bits into a single coarse-grained bin as in Example 1, we observe that to do this, the forecaster would likely encounter a large $\mathrm{calerr}(t)$ in the middle of the time horizon. For instance, suppose that the forecaster predicts $\frac{k+1}{2k}$ at every time step in Example 1. Then, the calibration error $\mathrm{calerr}(t)$ reduces to $|\Delta_p(t)| = |m_p(t) - n_p(t)\cdot p|$ for $p = \frac{k+1}{2k}$. Since each of the first $T/4$ bits has an expectation of at most $1/4$, the expected sum of these bits, $\mathbb{E}\left[m_p(T/4)\right]$, is at most $T/16$. On the other hand, $n_p(T/4)\cdot p = pT/4 \ge T/8$. It follows that $\mathrm{calerr}(T/4)$ will be as large as $\Omega(T)$ in expectation. Then, if the adversary deviates from the above construction and keeps outputting zeros in the remaining $3T/4$ time steps, $\mathrm{calerr}(T)$ will also be large.

This observation motivates the following "early stopping" trick: instead of directly lower bounding $\mathrm{calerr}(T)$, it suffices to show that $\mathrm{calerr}(t)$ is large at some step $t \in [T]$. Formally, define $\mathrm{maxerr}(t) := \max_{t' \in [t]} \mathrm{calerr}(t')$ as the maximum cumulative error encountered during the first t steps. The following proposition states that a high-probability lower bound on $\mathrm{maxerr}(T)$ implies the existence of another scheme that gives a high-probability lower bound on $\mathrm{calerr}(T)$.

Proposition 2. *Suppose that for $B, p > 0$, there exists a scheme $\mathcal{A}$ of the adversary that spans at most T time steps such that the inequality $\Pr\left[\mathrm{maxerr}(T^{\mathrm{actual}}) \ge B\right] \ge p$ holds for any forecaster, where random variable T^{actual} denotes the number of steps that $\mathcal{A}$ actually lasts. Then, there also exists a scheme that lasts exactly T time steps such that $\Pr[\mathrm{calerr}(T) \ge B/2] \ge p$ for any forecaster.*

Proof. We define another scheme $\mathcal{A}'$ that simulates $\mathcal{A}$. As soon as $\mathrm{calerr}(t_0) \ge B$ holds at some time t_0, $\mathcal{A}'$ deviates from $\mathcal{A}$ and computes $\sum_{p\in P}\Delta_p^+(t_0)$ and $\sum_{p\in P}\Delta_p^-(t_0)$. Since the two terms sum up to $\mathrm{calerr}(t_0)$, at least one of the terms is at least $B/2$. If $\sum_{p\in P}\Delta_p^+(t_0) \ge B/2$, scheme $\mathcal{A}'$ keeps outputting 1 in the remaining $T - t_0$ time steps; otherwise $\mathcal{A}'$ keeps outputting 0.

In the former case, $\sum_{p\in P}\Delta_p^+$ will never drop below $\sum_{p\in P}\Delta_p^+(t_0)$, so we have

$$\mathrm{calerr}(T) \ge \sum_{p\in P}\Delta_p^+(T) \ge \sum_{p\in P}\Delta_p^+(t_0) \ge B/2;$$

similarly, $\mathrm{calerr}(T) \ge \sum_{p\in P}\Delta_p^-(T) \ge \sum_{p\in P}\Delta_p^-(t_0) \ge B/2$ holds in the latter case. This shows that $\mathrm{maxerr}(T^{\mathrm{actual}}) \ge B$ when running scheme $\mathcal{A}$ implies that $\mathrm{calerr}(T) \ge B/2$ when running scheme $\mathcal{A}'$, and thus proves the proposition. □

Sidestepping. To prevent the forecaster from covering up the mistakes in the past, we note that such cover-ups are only possible if the probabilities of the later bits are in the "right direction" compared to the signs of Δ_p's. More concretely, in Example 2, it is crucial that the last $2T/3$ bits contain both zeros and ones for the cover-up to be possible. In contrast, if $\Delta_{1/2}(T/3) > 0$ and the remaining bits are all ones, predicting $1/2$ will only further increase $\Delta_{1/2}$ and thus increase the calibration error.

This motivates us to choose the probabilities in a *sidestepping* way, so that the error incurred by previous predictions cannot be fixed in the future. Suppose the adversary first flips a few fair coins with probability $1/2$. Then, assuming that all of the forecaster's

predictions are exactly $1/2$, the adversary switches to another probability based on the sign of $\Delta_{1/2}(t)$. If $\Delta_{1/2}(t) > 0$, the adversary switches to a coin with a larger bias $3/4$, so that if the forecaster keeps predicting $1/2$, $\Delta_{1/2}$ will only further increase in expectation; otherwise, if $\Delta_{1/2} < 0$, the bias is changed to $1/4$ accordingly. Similarly, after tossing the coin with probability $3/4$ for a while, the adversary changes the probability to either $5/8$ or $7/8$ depending on the sign of $\Delta_{3/4}$.

We could repeat the above procedure and choose the probabilities such that cover-ups are not possible. However, as soon as we change the probability of the bit $\Theta(\log T)$ times, all the valid choices of the probability would fall into an interval of length $1/T$, at which point the forecaster can afford to keep predicting the same probability later on, since rounding the predictions to the nearest multiple of $1/T$ only increases the calibration error by an additive $O(1)$ amount. Thus, applying the above scheme verbatim could only force the forecaster into predicting at most $k = O(\log T)$ different values, each corresponding to an epoch with T/k steps. Consequently, the resulting lower bound will be at best $\Omega(\sqrt{T \log T})$, which is not significantly better than the trivial bound.

Nevertheless, the actual construction of the scheme uses a similar strategy based on the idea of sidestepping. The key difference is that, instead of ensuring that the error incurred in *every* epoch cannot be covered up in later epochs, the actual construction only guarantees this for a substantial fraction of the epochs, which also turns out to be sufficient for proving the lower bound.

1.4 Related Work

The notion of calibration in the prediction context dates back to at least the 1950s. In the literature of meteorology, Brier [1] suggested that the quality of weather forecasts should be evaluated by comparing the forecast probability of rain and the actual fraction of rainy days among the days on which the probability is predicted. Calibration was later studied by Dawid [2] from a Bayesian perspective.

Foster and Vohra [6] studied the online prediction of arbitrary binary sequences from the calibration perspective, and proved the existence of a forecaster that is well-calibrated on any arbitrary binary sequence. While the results in the paper were stated in the asymptotic regime where T tends to infinity, the minimax proof due to Sergiu Hart ([6, Section 4] and [9]) directly implies an $O(T^{2/3})$ upper bound on the calibration error defined in (1). The work of Foster and Vohra was later simplified by [3, 7] and extended to settings where the calibration condition is tested on different subsets of the time horizon [5, 15, 18]. Vovk [20] further developed this approach and obtained non-asymptotic results.

The notion of calibration has also received increasing attention in the machine learning literature; see, e.g., [8, 13, 14, 21]. In binary classification, a classifier that maps data points to values in $[0, 1]$ is called well-calibrated if, among the data points on which value p is predicted, the fraction of positive examples is close to p. In other words, the outputs of well-calibrated classifiers can be interpreted as the probability that the data points belong to the positive class. One reason for the increased attention on calibration is that trained neural networks typically yield very poorly calibrated models. This classification setting is different from the online setting studied in this work, since the classifier makes the predictions for all the data points in a single batch. Thus, unlike the discussion in Example 2, it is impossible to cover up the error incurred on certain data points by strategically adjusting the remaining predictions.

Calibration has also been recently studied in the setting of algorithmic fairness [10–12, 16, 17, 19]. In this context, a predictor labels each individual from the population with a value in $[0, 1]$, which is intended to be the probability that the individual belongs to a specific class of interest. The calibration criterion proposed by [12] requires the predictions to be calibrated on a specified family of subsets of the population. When each subset in the family denotes a protected subset of the population, the calibration constraint prevents predictors that are discriminatory across different groups. [10] introduced another related fairness notion, *multicalibration*, which requires the predictions to be well-calibrated on every subgroup of the population that can be identified computationally.

1.5 Organization of the Paper

In the remainder of the paper, we first take a detour and introduce a two-player game called Sign-Preservation in Section 2. The Sign-Preservation game serves as a simplified abstraction of the sidestepping technique described in Section 1.3. We will state a reduction theorem (Theorem 4) in Section 2 and apply it to derive Theorem 1.

The rest of the paper is devoted to the proof of Theorem 4. In Section 3, we sketch the *sidestepping scheme* of the adversary as well as an idealized analysis of the scheme. We discuss a few challenges towards pinning down the optimal calibration error following our approach along with some other open problems in Section 4. Finally, in Section 5, we present the scheme formally and then prove Theorem 4.

2 DETOUR: THE SIGN-PRESERVATION GAME

As a detour, we introduce the following two-player sequential game called Sign-Preservation. We name these two players "player A" and "player F" to emphasize that they are analogous to the adversary and the forecaster in the prediction setting. An instance of the Sign-Preservation game with parameters k and r, denoted by Sign-Preservation(k, r), proceeds as follows. At the beginning of the game, there are k empty cells numbered $1, 2, \ldots, k$. The game consists of at most r rounds, and in each round:

(1) Player A may terminate the game immediately.
(2) Otherwise, player A chooses an empty cell with number $j \in [k]$.
(3) After knowing the value of j, player F places a *sign* (either "+" or "−") into cell j, and cell j is no longer empty.

When the game ends, we examine the signs placed by player F. We say that the sign in cell j is *removed* if either one of the following two holds:

- The sign is "+", and there exists $j' < j$ such that another sign is put into cell j' after this sign is put into cell j.
- The sign is "−", and there exists $j' > j$ such that another sign is put into cell j' after this sign is put into cell j.

If neither condition holds, the sign is said to be *preserved*. Equivalently, a "+" sign (resp. "−" sign) is preserved if and only if all

the subsequent signs are placed in cells with larger (resp. smaller) numbers. Player A's goal is to maximize the number of preserved signs, while player F tries to minimize this number.

2.1 Connection to Binary Prediction

Define the game value $\mathrm{opt}(k, r)$ as the maximum number of preserved signs in Sign-Preservation(k, r), assuming that both players play optimally. We call a pair of numbers (α, β) *admissible* if $\mathrm{opt}(k, k^\alpha)$ is lower bounded by $\Omega(k^\beta)$.

Definition 3. *$(\alpha, \beta) \in (0, 1]^2$ is admissible if there exists constant $c_0 > 0$ such that $\mathrm{opt}(k, r) \geq c_0 k^\beta$ holds for all integers $k \geq 1$ and $r \geq k^\alpha$.*

The following reduction theorem connects the Sign-Preservation game to the binary prediction setting. We will sketch the proof of Theorem 4 in Section 3 and present the full proof in Section 5.

Theorem 4. *Suppose that (α, β) is an admissible pair. Let $c = \frac{2\beta+1}{\alpha+2\beta+2}$. There exists a scheme of the adversary such that every forecaster incurs an expected calibration error of*

$$\mathbb{E}\left[\mathrm{calerr}(T)\right] \geq \Omega(T^c/\sqrt{\log T}).$$

Note that the exponent $c = \frac{2\beta+1}{\alpha+2\beta+2}$ is strictly greater than $1/2$ if and only if $\beta > \alpha/2$. In the remainder of this section, we will prove the existence of such an admissible pair (α, β) with $\beta > \alpha/2$ and then use it to prove Theorem 1.

2.2 Lower Bounding the Game Value

The following lemma gives two lower bounds on the optimal game value $\mathrm{opt}(\cdot, \cdot)$. The first states that Player A could make all signs preserved on Sign-Preservation$(k, \log k)$. The second states a "tensorization" property of the game, which allows us to lower bound a series of $\mathrm{opt}(\cdot, \cdot)$ given $\mathrm{opt}(k, r)$ for some specific k and r.

Lemma 5. *For any integer $t \geq 1$,*

(1) $\mathrm{opt}(2^t - 1, t) = t$;

(2) $\mathrm{opt}(a, b) \geq c \geq 1$ *implies* $\mathrm{opt}(a^t, b^t) \geq \left(\frac{c+1}{2}\right)^t$.

Another useful fact is the monotonicity of $\mathrm{opt}(k, r)$ in both k and r.

Lemma 6. *For any $1 \leq k_1 \leq k_2$ and $1 \leq r_1 \leq r_2$, $\mathrm{opt}(k_1, r_1) \leq \mathrm{opt}(k_2, r_2)$.*

Lemmas 5 and 6 are proved in Appendix A.

2.3 Proof of Theorem 1

The first part of Lemma 5 alone does not give any admissible pairs, because on an instance with k cells, the number of preserved signs is at most $O(\log k) = o(k^\beta)$ for any $\beta > 0$. However, when combined with the second part of Lemma 5, it indeed yields a non-trivial admissible pair, which in turn proves the lower bound in Theorem 1.

Proof of Theorem 1. Applying the first part of Lemma 5 with $t = 8$ gives $\mathrm{opt}(255, 8) = 8$. Then, the second part of Lemma 5, along with the trivial case $\mathrm{opt}(1, 1) = 1$, implies that $\mathrm{opt}(255^t, 8^t) \geq (9/2)^t$ for any integer $t \geq 0$. Let $\alpha = \frac{\log 8}{\log 255}$ and $\beta = \frac{\log(9/2)}{\log 255}$. We will prove in the following that (α, β) is admissible. Then, Theorem 1 would directly follow from Theorem 4.

Fix $k \geq 1$, $r \geq k^\alpha$, and let $t = \lfloor \frac{\log k}{\log 255} \rfloor$. We have $k \geq 255^t$ and $t > \frac{\log k}{\log 255} - 1$. Furthermore, $r \geq k^\alpha \geq 255^{\alpha t} = 8^t$. By Lemma 6, we have

$$\mathrm{opt}(k, r) \geq \mathrm{opt}(255^t, 8^t) \geq (9/2)^t > (9/2)^{\frac{\log k}{\log 255} - 1} = \frac{2}{9} k^\beta.$$

This shows that (α, β) is an admissible pair, and thus proves the theorem. □

3 OVERVIEW OF THE PROOF

In this section, we sketch a simplified version of the *sidestepping scheme*, which will be used to prove Theorem 4. We will then explain how the Sign-Preservation game captures the essence of the scheme by drawing an analogy between the game and the sidestepping scheme. Finally, we present an idealized analysis that contains most of the key ideas behind the formal proof in Section 5.

3.1 A Sketch of the Scheme

The sidestepping scheme is based on the notion of *epochs*. The time horizon $1, 2, \ldots, T$ is divided into k epochs of the same length T/k. The scheme chooses a probability p_i^* at the beginning of the i-th epoch, and outputs T/k independent samples from $\mathrm{Ber}(p_i^*)$ during this epoch. In the ideal case, we expect the forecaster to keep predicting a probability $\approx p_i^*$ throughout epoch i. Then, we would be able to lower bound $\mathrm{calerr}(T)$ by $k \cdot \sqrt{T/k} = \sqrt{Tk}$.

As discussed in Section 1.2, this straightforward construction is vulnerable to untruthful forecasters whose predictions can be far away from the actual probability p_i^*. In particular, we need to prevent the forecaster from: (1) merging different epochs into a larger, coarse-grained bin, i.e., by predicting the average of p_i^* at every time step; (2) covering up the errors made in the previous epochs. Resolving the first issue is relatively easier. Suppose that we choose the probabilities p_1^* through p_k^* to be $1/k, 2/k, \ldots, k/k$ and, in some epoch i, a significant fraction of the predictions are $(1/k)$-far from the actual probability p_i^*. Since each epoch has length T/k, these predictions lead to a calibration error of $(T/k) \cdot (1/k) = T/k^2$ in expectation. Then, we will be able to catch this error using the early stopping technique (Proposition 2).

Otherwise, suppose that most of the forecaster's predictions are $(1/k)$-close to the true probability p_i^* over epoch i. For simplicity, we assume for now that all the T/k predictions are exactly p_i^*. Then, standard tail bounds for the binomial distribution imply that we expect an error of $|\Delta_{p_i^*}| \geq \Omega(\sqrt{T/k})$ after epoch i. Thus, summing over the k epochs and taking a minimum with T/k^2 (the error when the forecaster is untruthful) seems to suggest an lower bound of $\min(T/k^2, \sqrt{Tk})$.

The issue with the above argument is that the forecaster might be able to cover up its error in later epochs, unless the probabilities of the future epochs are all in the right direction. For example, if $\Delta_{p_i^*} > 0$ at the end of epoch i, and the probability chosen for the next epoch satisfies $p_{i+1}^* < p_i^*$, the forecaster can decrease $\Delta_{p_i^*}$ by keeping predicting p_i^* in epoch $i + 1$, until $\Delta_{p_i^*}$ becomes close to zero. Fortunately, this kind of "cover-up" would not be

possible if we chose the probabilities such that $p^*_{i'} > p^*_i$ for every $i' = i+1, i+2, \ldots, k$. This observation motivates us to choose $p^*_1, p^*_2, \ldots, p^*_k$ more carefully, so that the number of epochs whose $|\Delta_{p^*_i}|$ are preserved at the end of the scheme is maximized.

3.2 Analogy between Binary Prediction and Sign-Preservation

The above discussion closely resembles the Sign-Preservation game that we defined. In the scheme sketched above, we had k possible choices, $1/k$ through k/k, for each p^*_i, and the j-th largest probability j/k corresponds to the j-th cell in Sign-Preservation. The i-th epoch of the scheme is modeled by the i-th round of the game: (1) player A's action of choosing cell j corresponds to the adversary's choice of $p^*_i = j/k$ for epoch i; (2) player F's action of placing a "+"/"−" sign can be thought of as getting $\Delta_{p^*_i} > 0$ or $\Delta_{p^*_i} < 0$ at the end of epoch i. Finally, a sign gets removed by another sign placed later (if the other sign is on the proper direction), since the error in epoch i could be fixed by a later epoch i', given that the sign of $p^*_{i'} - p^*_i$ is opposite to that of $\Delta_{p^*_i}$. Consequently, the number of epochs whose $|\Delta_{p^*_i}|$ are retained at the end of the scheme is modeled by the number of preserved signs at the end of the Sign-Preservation game.

3.3 Proof Sketch of Theorem 4

We sketch a proof of Theorem 4 in the following. Let (α, β) be an admissible pair, and k be a parameter to be determined later. In contrast to the scheme described above, we will divide the time horizon into k^α epochs instead, and each epoch has length T/k^α.

The adversary simulates an instance of Sign-Preservation(k, k^α) where player A plays optimally. Every time player A chooses a cell with number j, the adversary chooses $p^*_i = j/k$ for the next epoch i. In other words, the next T/k^α bits will be independent samples from $\mathrm{Ber}(j/k)$.

Within epoch i, we say that a prediction given by the forecaster is *untruthful* if the predicted probability is $(1/k)$-far from p^*_i. Clearly, the forecaster has to make either $\Omega(T/k^\alpha)$ untruthful predictions, or $\Omega(T/k^\alpha)$ truthful ones. In the former case, we claim that each untruthful prediction increases the cumulative calibration error by an $\Omega(1/k)$ amount, so the total increase throughout this epoch will be at least $\Omega(T/k^{\alpha+1})$. Then, the adversary would be able to catch this $\Omega(T/k^{\alpha+1})$ error using the "early stopping" trick.

Otherwise, suppose that epoch i is truthful. Then, lower bounds on binomial tails imply that there will be an $\Omega(\sqrt{T/k^\alpha})$ error in expectation after this epoch. Furthermore, this error cannot be significantly reduced if no later epochs is assigned a probability $p^*_{i'}$ with $(p^*_{i'} - p^*_i) \cdot \Delta_{p^*_i} < 0$. Thus, if we choose (on behalf of player F in Sign-Preservation) the sign for this cell as the sign of $\Delta_{p^*_i}$, $|\Delta_{p^*_i}|$ will still be $\Omega(\sqrt{T/k^\alpha})$ at the end of the scheme as long as the sign placed in this round is preserved at the end of the Sign-Preservation game.

Since (α, β) is admissible, there will be at least $\Omega(k^\beta)$ preserved signs, thus giving a lower bound of $\Omega(k^\beta \sqrt{T/k^\alpha})$. Taking a minimum with the $\Omega(T/k^{\alpha+1})$ error (in case of an epoch with too many untruthful predictions) and plugging in the optimal choice of $k = T^{\frac{1}{\alpha+2\beta+2}}$ gives the claimed $\Omega(T^c)$ lower bound for $c = \frac{2\beta+1}{\alpha+2\beta+2}$.

4 DISCUSSION

In this section, we discuss a few open directions for further understanding the optimal calibration error that can be achieved in the binary prediction setting.

Gap between upper and lower bounds. In light of Theorem 4, an immediate open problem is to find other admissible pairs (α, β) that result in a larger exponent $c = \frac{2\beta+1}{\alpha+2\beta+2}$ in the lower bound. In particular, the best possible exponent we can get from Theorem 4 is $c = 3/5$ if $(1, 1)$ is admissible, i.e., $\Omega(k)$ signs can be preserved in a Sign-Preservation game with k cells and k rounds. Either proving or disproving this would help us to understand the limit of the approach based on the sidestepping scheme and Theorem 4.

Another natural open question is whether the $O(T^{2/3})$ upper bound is indeed optimal. In particular, can we construct a better forecaster by proving a converse of Theorem 4 that translates upper bounds on $\mathrm{opt}(\cdot, \cdot)$ into actual strategies for the forecaster? While such a converse is likely to exist when the adversary is epoch-based (and even announces the probability of each epoch at the beginning of that epoch), extending this reduction in the converse direction to more general cases seems challenging.

The power of adaptivity. Our proof of the lower bound is based on an adaptive scheme for the adversary. More exactly, the scheme uses adaptivity on two different levels: (1) The adversary decides the probability p^*_i of an epoch i based on the gameplay of a Sign-Preservation instance. In general, player A of Sign-Preservation is allowed to choose the cells adaptively based on the state of the game, which in turn means that p^*_i are chosen adaptively; (2) When we formally prove Theorem 4, we will need to apply Proposition 2 to transform the scheme—which is only guaranteed to give a large $\mathrm{calerr}(t)$ at some point t—into another scheme with a large $\mathbb{E}[\mathrm{calerr}(T)]$, and the transformation based on the early stopping trick is inherently adaptive.

Nevertheless, we conjecture that both uses of adaptivity can be replaced by randomization: (1) The lower bounds on $\mathrm{opt}(\cdot, \cdot)$ in Lemma 5 can still be achieved (up to a constant factor) in expectation by a non-adaptive yet randomized strategy for player A; (2) The adaptive early stopping strategy in the proof of Proposition 2 can also be replaced by a randomized one, e.g., that chooses the number of epochs uniformly at random from $1, 2, \ldots, k^\alpha$. Thus, as far as super-$\sqrt{T}$ lower bounds are concerned, adaptivity appears inessential to the adversary, though adaptivity does greatly simplify the analysis of the scheme. Furthermore, it remains an interesting yet challenging open problem to understand whether the extra power brought by the adaptivity increases the calibration error that the optimal forecaster has to incur.

5 PROOF OF THEOREM 4

5.1 The Sidestepping Scheme

We formally define the sidestepping scheme in Algorithm 1 and the epochs in the scheme are defined in Algorithm 2. The core of the scheme is to simulate an instance of Sign-Preservation(k, k^α) for some carefully chosen k. In this simulated game, player A plays the optimal strategy while the sidestepping scheme, perhaps paradoxically, plays on behalf of player F. This situation can be best

illustrated as Figure 1, where the sidestepping scheme plays as a "man-in-the-middle" and connects an optimal player A for Sign-Preservation to the forecaster in the prediction setting.

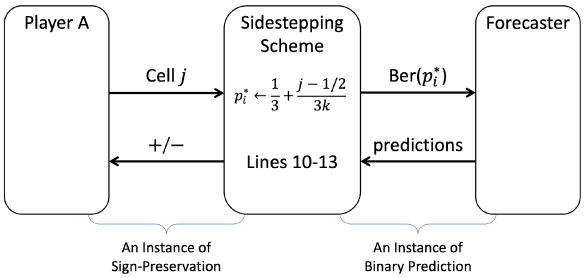

Figure 1: The sidestepping scheme working as a man-in-the-middle. Player A is playing Sign-Preservation and the forecaster is in the binary prediction setting from their perspectives.

Algorithm 1: Sidestepping Scheme

Input: Horizon length T and parameters $\alpha, \beta \in (0, 1]$.

$k \leftarrow T^{1/(\alpha+2\beta+2)}$; $\theta \leftarrow \frac{1}{1440}\sqrt{\frac{T}{k^\alpha \ln T}}$;
Simulate an instance of Sign-Preservation(k, k^α);
for $i = 1, 2, \ldots, k^\alpha$ **do**
  if *player A terminates the game in round i* **then**
    break;
  Let $j \in [k]$ be the cell chosen by player A in round i;
  $\mathcal{I}_i \leftarrow$ interval $(\frac{1}{3} + \frac{j-1}{3k}, \frac{1}{3} + \frac{j}{3k})$;
  $p_i^* \leftarrow \frac{1}{3} + \frac{j-1/2}{3k}$; `// the middle point of` $\mathcal{I}_i$
  Call Epoch$(T/k^\alpha, \mathcal{I}_i, p_i^*, \theta)$;
  if $\sum_{p \in P \cap \mathcal{I}_i} \Delta_p^+ \geq \sum_{p \in P \cap \mathcal{I}_i} \Delta_p^-$ **then**
    Let player F place "+" into cell j;
  else
    Let player F place "−" into cell j;
end

Algorithm 2: Epoch$(m, \mathcal{I}, p^*, \theta)$

for $i = 1, 2, \ldots, m$ **do**
  if $\sum_{p \in P \cap \mathcal{I}} |\Delta_p| \geq \theta$ **then**
    break;
  Draw $b \sim \text{Ber}(p^*)$;
  Output bit b;
end

The scheme differs from the simplified version in Section 3 in the following two aspects. First, the probability p_i^* is restricted to the interval $[1/3, 2/3]$ (instead of $[0, 1]$), so that the binomial distribution with parameter p_i^* would have a tail that is lower bounded by Gaussian tails. More specifically, when player A chooses some cell j in the game, we start an epoch associated with probability $p_i^* = \frac{1}{3} + \frac{j-1/2}{3k}$. Note that p_i^* is exactly the middle point of the j-th interval when $[1/3, 2/3]$ is partitioned into k intervals of length $1/(3k)$.

Second, an epoch may span less than T/k^α time steps. In particular, we set a threshold θ and end an epoch as soon as the interval $\mathcal{I}$ associated with the epoch already contributes at least θ to the cumulative error. The purpose of this slight change is mostly to simplify the analysis. As a result, the sidestepping scheme may end before T time steps. In the following, we use random variable T^{actual} to denote the number of time steps that the scheme actually lasts. At the end of the proof, we will transform the sidestepping scheme into another scheme that spans exactly T steps using Proposition 2.

5.2 Classification of Epochs

For each possible execution of the sidestepping scheme (Algorithm 1), we say that an epoch is *untruthful* if the forecaster makes too many predictions that are $\Omega(1/k)$-far away from the actual probability; otherwise it is said to be *truthful*.

Definition 7 (Untruthful epochs). *An epoch i associated with interval $\mathcal{I}_i$ is untruthful if, within epoch i, the forecaster makes at least $T/(2k^\alpha)$ predictions with values outside $\mathcal{I}_i$.*

We call a truthful epoch *negligible* if, when the epoch ends, the interval associated with it contributes less than $\theta = \frac{1}{1440}\sqrt{\frac{T}{k^\alpha \ln T}}$ to the cumulative calibration error at that time; otherwise the epoch is said to be *non-negligible*. By our definition of Epoch (Algorithm 2), an epoch is negligible only if it takes exactly m time steps.

Definition 8 (Negligible epochs). *A truthful epoch i associated with interval $\mathcal{I}_i$ is negligible if, when epoch i ends at time step t, it holds that $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p(t)| < \theta$.*

Finally, for a truthful and non-negligible epoch, we call it *covered* if, at the end of the scheme, its contribution to calerr(T^{actual}) is less than $\theta/4$; otherwise we call it *uncovered*. In other words, an epoch is covered if the predictions in later epochs cover up a significant portion of the mistakes made by the forecaster in the epoch.

Definition 9 (Covered epochs). *A truthful and non-negligible epoch i associated with interval $\mathcal{I}_i$ is covered if, when the scheme ends after T^{actual} time steps, it holds that $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p(T^{\text{actual}})| < \theta/4$.*

Pictorially, the relation between different classes of epochs defined in Definitions 7 through 9 is demonstrated in Figure 2.

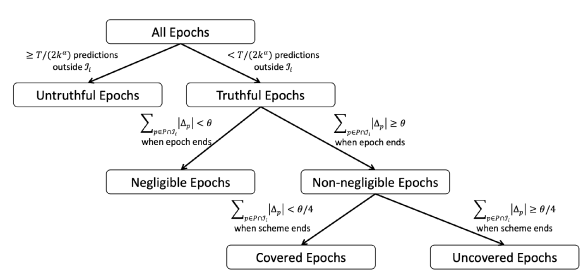

Figure 2: The relation between different classes of epochs.

5.3 Auxiliary Lemmas

Lemmas 10, 11 and 12 state that the following three hold with high probability: (1) either every epoch is truthful, or $\mathrm{calerr}(t)$ is large at some point t; (2) every truthful epoch is non-negligible; (3) for every non-negligible epoch i, if the sign placed in round i of Sign-Preservation is preserved, epoch i is uncovered.

Lemma 10. *Fix $i \in [k^\alpha]$ and let $B := \frac{T}{48k^{\alpha+1}}$. Suppose that epoch i spans the time steps $t_0+1, t_0+2, \ldots, t_1$. The probability that epoch i is untruthful and $\mathrm{calerr}(t) < B$ holds for every $t = t_0, t_0+1, \ldots, t_1$ is at most $\exp(-\Omega(T/k^{\alpha+2})) = o(1/T)$.*

Lemma 11. *For any fixed $i \in [k^\alpha]$, the probability that epoch i is truthful and negligible is at most $T^{-2} = o(1/T)$.*

Lemma 12. *For any fixed $i \in [k^\alpha]$, the probability that the following two hold simultaneously is at most $T \cdot \exp(-\theta/(12k)) = o(1/T)$: (1) epoch i is truthful, non-negligible, and covered; (2) the sign placed in the i-th round of the Sign-Preservation game is preserved.*

All these three lemmas are proved by applying standard concentration and anti-concentration bounds to carefully chosen quantities tailored to the epoch in question. The proofs are deferred to Appendix B.

5.4 Putting Everything Together

Now we are ready to prove Theorem 4.

Proof of Theorem 4. Let $k = T^{1/(\alpha+2\beta+2)}$, $\theta = \frac{1}{1440}\sqrt{\frac{T}{k^\alpha \ln T}}$ as in Algorithm 1, and $c = \frac{2\beta+1}{\alpha+2\beta+2}$ as in the statement of the theorem. Define $B := \min\left(T/(48k^{\alpha+1}), c_0\theta k^\beta/4\right) = \Omega\left(T^c/\sqrt{\log T}\right)$, where c_0 is the constant for the admissible pair (α, β) in Definition 3. Recall that $T^{\mathrm{actual}} \le T$ denotes the number of time steps that the sidestepping scheme actually lasts, and $\mathrm{maxerr}(T^{\mathrm{actual}})$ denotes $\max_{t \in [T^{\mathrm{actual}}]} \mathrm{calerr}(t)$. We will show that the sidestepping scheme defined as Algorithm 1, when running against any forecaster, satisfies that $\Pr\left[\mathrm{maxerr}(T^{\mathrm{actual}}) \ge B\right] \ge 1 - o(1)$.

To upper bound $\Pr\left[\mathrm{maxerr}(T^{\mathrm{actual}}) < B\right]$, we define $\mathcal{E}^{\mathrm{truth}}$ as the event that all the k^α epochs are truthful (in the sense of Definition 7). Then, we note that

$$\begin{aligned}
&\Pr\left[\mathrm{maxerr}(T^{\mathrm{actual}}) < B\right] \\
\le &\sum_{i=1}^{k^\alpha} \Pr\left[\text{epoch } i \text{ is untruthful} \wedge \mathrm{maxerr}(T^{\mathrm{actual}}) < B\right] \\
&+ \Pr\left[\mathcal{E}^{\mathrm{truth}} \wedge \mathrm{maxerr}(T^{\mathrm{actual}}) < B\right] \qquad \text{(union bound)} \\
\le &\sum_{i=1}^{k^\alpha} \Pr\left[\text{epoch } i \text{ is untruthful} \wedge \mathrm{maxerr}(T^{\mathrm{actual}}) < B\right] \\
&+ \Pr\left[\mathcal{E}^{\mathrm{truth}} \wedge \mathrm{calerr}(T^{\mathrm{actual}}) < B\right]. \\
&\qquad (\mathrm{maxerr}(T^{\mathrm{actual}}) < B \implies \mathrm{calerr}(T^{\mathrm{actual}}) < B)
\end{aligned}$$

By Lemma 10, each term in the summation is at most $o(1/T)$, so the whole summation is upper bounded by $k^\alpha \cdot o(1/T) = o(1)$. It remains to prove $\Pr\left[\mathcal{E}^{\mathrm{truth}} \wedge \mathrm{calerr}(T^{\mathrm{actual}}) < B\right] = o(1)$.

Let $\mathcal{E}^{\mathrm{negl}}$ denote the event that at least one of the k^α epochs is truthful and negligible, and $\mathcal{E}^{\mathrm{cover}}$ be the event that there exists $i \in [k^\alpha]$ such that: (1) epoch i is truthful, non-negligible and covered; (2) the sign placed in the i-th round of Sign-Preservation is preserved in the end. Then, by Lemmas 11 and 12 and a union bound over the $k^\alpha \le T$ epochs, $\Pr\left[\mathcal{E}^{\mathrm{truth}} \wedge (\mathcal{E}^{\mathrm{negl}} \vee \mathcal{E}^{\mathrm{cover}})\right] = o(1)$. We will show in the following that the event $\mathcal{E}^{\mathrm{truth}} \wedge \mathrm{calerr}(T^{\mathrm{actual}}) < B$ is a subset of $\mathcal{E}^{\mathrm{truth}} \wedge (\mathcal{E}^{\mathrm{negl}} \vee \mathcal{E}^{\mathrm{cover}})$, and thus it holds that

$$\begin{aligned}
&\Pr\left[\mathcal{E}^{\mathrm{truth}} \wedge \mathrm{calerr}(T^{\mathrm{actual}}) < B\right] \\
\le &\Pr\left[\mathcal{E}^{\mathrm{truth}} \wedge (\mathcal{E}^{\mathrm{negl}} \vee \mathcal{E}^{\mathrm{cover}})\right] = o(1).
\end{aligned}$$

We will prove the contrapositive: assuming that event $\mathcal{E}^{\mathrm{truth}}$ happens yet neither $\mathcal{E}^{\mathrm{negl}}$ nor $\mathcal{E}^{\mathrm{cover}}$ happens, $\mathrm{calerr}(T^{\mathrm{actual}}) \ge B$ holds. Let $S \subseteq [k^\alpha]$ denote the set of indices i such that the sign placed in round i of the Sign-Preservation game is preserved. Since (α, β) is admissible with constant c_0, $|S| \ge c_0 k^\beta$. For each $i \in S$, since we assumed $\mathcal{E}^{\mathrm{truth}} \wedge \overline{\mathcal{E}^{\mathrm{negl}}} \wedge \overline{\mathcal{E}^{\mathrm{cover}}}$, epoch i is truthful, non-negligible and uncovered. Then, by definition, it holds that $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p(T^{\mathrm{actual}})| \ge \theta/4$. Since the intervals $\mathcal{I}_i$ are disjoint for different indices $i \in S$, we have

$$\mathrm{calerr}(T^{\mathrm{actual}}) \ge \sum_{i \in S} \sum_{p \in P \cap \mathcal{I}_i} |\Delta_p(T^{\mathrm{actual}})| \ge c_0 k^\beta \cdot (\theta/4) \ge B.$$

This completes the proof of $\Pr\left[\mathrm{maxerr}(T^{\mathrm{actual}}) \ge B\right] \ge 1 - o(1)$ when running Algorithm 1 against any forecaster. For sufficiently large T, the $o(1)$ term is at most $1/2$. Then, by Proposition 2, there exists a scheme such that $\Pr[\mathrm{calerr}(T) \ge B/2] \ge 1/2$, which implies the lower bound $\mathbb{E}[\mathrm{calerr}(T)] \ge B/4 = \tilde{\Omega}(T^c)$. □

ACKNOWLEDGMENTS

We would like to thank Dean P. Foster for bringing to our attention this calibration perspective on online predictions as well as the problem of proving super-$\sqrt{T}$ calibration lower bounds, and for his comments and suggestions on an earlier draft of this paper. This work was supported by NSF Awards CCF-1704417 and AF-1813049, DOE Award DE-SC0019205, and ONR Young Investigator Award N00014-18-1-2295.

A DEFERRED PROOFS FROM SECTION 2

Proof of Lemma 5. The first part follows from a simple strategy that resembles a binary search: Player A chooses cell 2^{t-1} in the first round. If the sign placed by player F is "+", proceed with the remaining $t-1$ rounds on the $2^{t-1}-1$ cells with numbers $2^{t-1}+1, 2^{t-1}+2, \ldots, 2^t-1$; otherwise, proceed with cells $1, 2, \ldots, 2^{t-1}-1$. Then, none of the t signs will be removed in the end, and this proves $\mathrm{opt}(2^t-1, t) = t$.

We prove the second part by induction. The inequality clearly holds for $t = 1$ since the assumption implies $\mathrm{opt}(a, b) \ge c \ge (c+1)/2$. For $t \ge 2$, we consider the following strategy for the instance Sign-Preservation(a^t, b^t): Player A divides the a^t cells into a "super-cells", each corresponding to a^{t-1} contiguous cells. Then, Player A simulates a hypothetical instance of Sign-Preservation(a, b), denoted by $\mathsf{SP}^{\mathrm{outer}}$, in the following sense: When one of the a super-cells is chosen in round i of $\mathsf{SP}^{\mathrm{outer}}$, Player A simulates an actual instance of Sign-Preservation(a^{t-1}, b^{t-1}), denoted by $\mathsf{SP}_i^{\mathrm{inner}}$, on the a^{t-1} cells corresponding to that super-cell, i.e., whenever a cell is chosen in $\mathsf{SP}_i^{\mathrm{inner}}$, Player A chooses the corresponding cell in the

actual Sign-Preservation(a^t, b^t) instance. After $\mathsf{SP}_i^{\text{inner}}$ terminates, Player A, on behalf of the "Player F" in $\mathsf{SP}^{\text{outer}}$, places a sign into that super-cell according to the majority of the preserved signs in $\mathsf{SP}_i^{\text{inner}}$.

Now we count the preserved signs in the original game. If the sign in a super-cell is preserved at the end of $\mathsf{SP}^{\text{outer}}$, any preserved sign in the a^{t-1} corresponding cells (at the end of the corresponding $\mathsf{SP}^{\text{inner}}$ instance) that agrees with the sign in that super-cell will also be preserved at the end of Sign-Preservation(a^t, b^t). By our choice of the sign's direction, there are at least $\text{opt}(a^{t-1}, b^{t-1})/2$ such signs for each preserved sign in $\mathsf{SP}^{\text{outer}}$. Moreover, for the preserved sign that is placed in the last round of $\mathsf{SP}^{\text{outer}}$, all the $\geq \text{opt}(a^{t-1}, b^{t-1})$ remaining signs in the corresponding super-cell will be preserved. By the inductive hypothesis that $\text{opt}(a^{t-1}, b^{t-1}) \geq \left(\frac{c+1}{2}\right)^{t-1}$, we have

$$\text{opt}(a^t, b^t) \geq (c-1)\cdot\frac{\text{opt}(a^{t-1}, b^{t-1})}{2} + \text{opt}(a^{t-1}, b^{t-1}) \geq \left(\frac{c+1}{2}\right)^t,$$

which completes the induction. □

Proof of Lemma 6. On an instance of Sign-Preservation(k_2, r_2), Player A can simulate the optimal strategy for the game instance Sign-Preservation(k_1, r_1) on cells $1, 2, \ldots, r_1$. Player A ends the game when the Sign-Preservation(k_1, r_1) instance terminates. Since $r_1 \leq r_2$, the simulated game never lasts more than r_2 rounds. By definition, there will be at least $\text{opt}(k_1, r_1)$ preserved signs, so we have $\text{opt}(k_2, r_2) \geq \text{opt}(k_1, r_1)$. □

B DEFERRED PROOFS FROM SECTION 5

The following concentration bound is an immediate corollary of the Azuma-Hoeffding inequality for submartingales.

Lemma 13. *Suppose that random variables $X_1, X_2, \ldots, X_m$ satisfy that for every $t \in [m]$: (1) $X_t \in [-1, 1]$ almost surely; (2) $\mathbb{E}[X_t | X_1, X_2, \ldots, X_{t-1}] \geq \mu$. Then, for any $c < m\mu$,*

$$\Pr\left[\sum_{t=1}^{m} X_t \leq c\right] \leq \exp\left(-\frac{(m\mu - c)^2}{2m}\right).$$

The following anti-concentration bound for binomial distributions follows from the Berry-Esseen theorem.

Lemma 14. *Suppose that $p \in [1/3, 2/3]$, Z follows the binomial distribution $B(m, p)$, and g follows the standard Gaussian distribution $N(0, 1)$. For any $c \in \mathbb{R}$, it holds that*

$$\left|\Pr\left[\frac{Z - mp}{\sqrt{mp(1-p)}} \geq c\right] - \Pr[g \geq c]\right| \leq O\left(\frac{1}{\sqrt{m}}\right),$$

where the $O(\cdot)$ notation hides a universal constant that does not depend on m or p.

Now we are ready to prove Lemmas 10 through 12.

Proof of Lemma 10. Let $\mathcal{I}_i = (l_i, r_i)$ be the interval associated with epoch i and $p_i^* = (l_i + r_i)/2$ be its middle point. By the choice of $\mathcal{I}_i$ and p_i^* in Algorithm 1, $p_i^* - l_i = r_i - p_i^* = \frac{1}{6k}$. Define

$$\hat{\Delta}(t) := \sum_{p \in P \cap [0, l_i]} \Delta_p(t) + \sum_{p \in P \cap [r_i, 1]} [-\Delta_p(t)]$$

as a proxy of calerr(t). It can be easily verified that both $\hat{\Delta}(t) \leq \text{calerr}(t)$ and $-\hat{\Delta}(t) \leq \text{calerr}(t)$.

Clearly, a prediction with value inside $\mathcal{I}_i$ does not change the value of $\hat{\Delta}$; in contrast, whenever a probability outside $\mathcal{I}_i$ is predicted, $\hat{\Delta}$ is incremented by at least $\Omega(1/k)$ in expectation. To see this, suppose that the forecaster predicts $p(t) \leq l_i$ at time step t. Then, the expected increment in $\hat{\Delta}$ is given by

$$\begin{aligned}
&\mathbb{E}\left[\Delta_{p(t)}(t) - \Delta_{p(t)}(t-1)\right] \\
=&\mathbb{E}\left[m_{p(t)}(t) - m_{p(t)}(t-1)\right] - p(t)\cdot\mathbb{E}\left[n_{p(t)}(t) - n_{p(t)}(t-1)\right] \\
=&p_i^* - p(t) \geq p_i^* - l_i = \frac{1}{6k}.
\end{aligned}$$

Moreover, the increment in $\hat{\Delta}$ is always bounded between -1 and 1. Similarly, whenever a prediction $p(t) \geq r_i$ is made, the increment in $\hat{\Delta}$ is always between -1 and 1 and has expectation $p(t) - p_i^* \geq r_i - p_i^* = \frac{1}{6k}$.

Let $m := T/(2k^\alpha)$. Assuming that epoch i is untruthful, there exists a unique time step $t_2 \in [t_0, t_1]$ when the forecaster makes the m-th prediction that falls outside $\mathcal{I}_i$. We will prove that $\hat{\Delta}(t_2) - \hat{\Delta}(t_0) \geq \frac{m}{12k}$ with high probability, which implies that either $\hat{\Delta}(t_2) \geq \frac{m}{24k}$ or $\hat{\Delta}(t_0) \leq -\frac{m}{24k}$. Then, we would have $\max(\text{calerr}(t_0), \text{calerr}(t_2)) \geq \frac{m}{24k} = B$ as desired. Indeed, our discussion above indicates that $\hat{\Delta}(t_2) - \hat{\Delta}(t_0)$ can be written as a sum of m random variables $X_1, X_2, \ldots, X_m$ satisfying that for each $j \in [m]$: (1) $X_j \in [-1, 1]$ almost surely; (2) $\mathbb{E}\left[X_j | X_1, X_2, \ldots, X_{j-1}\right] \geq \frac{1}{6k}$. Then, by the Azuma-Hoeffding inequality (in the form of Lemma 13), it holds that

$$\Pr\left[\hat{\Delta}(t_2) - \hat{\Delta}(t_0) \leq \frac{m}{12k}\right] \leq \exp\left(-\frac{m}{288k^2}\right) = \exp\left(-\Omega\left(\frac{T}{k^{\alpha+2}}\right)\right).$$

Finally, since Algorithm 1 chooses $k = T^{\frac{1}{\alpha+2\beta+2}}$ and requires $\beta > 0$, we have $\frac{T}{k^{\alpha+2}} = T^{\frac{2\beta}{\alpha+2\beta+2}} = \Omega(\text{poly}(T))$. Thus, $\exp(-\Omega(T/k^{\alpha+2})) = o(1/T)$, which completes the proof. □

Proof of Lemma 11. Let $\mathcal{I}_i = (l_i, r_i)$ be the interval associated with epoch i. For epoch i to be truthful but negligible, the forecaster needs to make at least $T/(2k^\alpha)$ predictions with values inside $\mathcal{I}_i$. Let $m = T/(324k^\alpha \ln T)$. We may further decompose epoch i into $162 \ln T$ blocks, each with at least m predictions that fall into $\mathcal{I}_i$. To prove the lemma, it suffices to show that, conditioning on the bits and predictions before each block, the probability that $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p|$ remains less than θ throughout the block is at most $1 - 3^{-4}$. Assuming this, the probability that epoch i becomes negligible after the $162 \ln T$ blocks is at most $(1 - 3^{-4})^{162 \ln T} \leq e^{-2 \ln T} = T^{-2}$, as claimed by the lemma.

Fix a block with $\geq m$ predictions inside $\mathcal{I}_i$ and let t_0 be the time step before the start of the block. Let $\delta := 1/(10\sqrt{m})$ and p_i^* be the middle point of $\mathcal{I}_i$. We divide the prediction values into the following four groups:

- $P_1 := \{p \in P \cap \mathcal{I}_i : \Delta_p(t_0) \geq 0, p_i^* - p \geq -\delta)\}$.
- $P_2 := \{p \in P \cap \mathcal{I}_i : \Delta_p(t_0) \geq 0, p_i^* - p < -\delta\}$.
- $P_3 := \{p \in P \cap \mathcal{I}_i : \Delta_p(t_0) < 0, p_i^* - p \leq \delta\}$.
- $P_4 := \{p \in P \cap \mathcal{I}_i : \Delta_p(t_0) < 0, p_i^* - p > \delta\}$.

Note that each prediction at probability p increases Δ_p by $p_i^* \cdot (1 - p) + (1 - p_i^*) \cdot (0 - p) = p_i^* - p$ in expectation. Thus, the above

definition basically says that each prediction that falls into $P_1 \cup P_3$ will either push Δ_p away from 0, or push it towards 0 by at most $\delta = O(1/\sqrt{m})$. In contrast, each prediction in $P_2 \cup P_4$ will push Δ_p in the opposite direction of the sign of $\Delta_p(t_0)$ by at least $\delta = \Omega(1/\sqrt{m})$.

Since there are m predictions inside $\mathcal{I}_i$ within the block, at least one of the four sets P_1 through P_4 receives at least $m/4$ predictions. In the remainder of the proof, we will prove the following claim: for each P_j, with probability at least 1/3, $\sum_{p \in P_j} |\Delta_p|$ will reach θ before or exactly when the forecaster makes the $(m/4)$-th prediction with value inside P_j. Thus, the scheme would have terminated before that. Assuming this, we may pretend that there are four independent bit sequences, each consisting of independent samples from $\mathrm{Ber}(p_i^*)$. When a probability inside P_j is predicted, the bit output by the adversary actually comes from the j-th sequence. Since the four bit sequences are independent, $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p|$ will reach θ at some point in this block with probability at least 3^{-4}. This would then prove the lemma.

Proofs for P_1 and P_3. The proofs for P_1 and P_3 are symmetric, so we only present the proof for P_1 in the following. Define the quantity $\hat{\Delta}(t) := \sum_{p \in P_1} \Delta_p(t)$. Whenever a value $p \in P_1$ is predicted, $\hat{\Delta}$ is incremented by $b - p$, where $b \sim \mathrm{Ber}(p^*)$. Moreover, by definition of P_1, every $p \in P_1$ is upper bounded by $p_i^* + \delta$. Thus, the increment of $\hat{\Delta}$ after the first $m/4$ predictions in P_1 is lower bounded by $Z - (p_i^* + \delta) \cdot (m/4)$, where Z follows the binomial distribution $B(m/4, p_i^*)$. Let $g \sim N(0,1)$ be a standard Gaussian random variable. By Lemma 14,

$$
\begin{aligned}
&\Pr\left[Z \geq mp_i^*/4 + m\delta/2\right] \\
&= \Pr\left[\frac{Z - mp_i^*/4}{\sqrt{mp_i^*(1-p_i^*)/4}} \geq \frac{m\delta/2}{\sqrt{mp_i^*(1-p_i^*)/4}}\right] \\
&\geq \Pr\left[\frac{Z - mp_i^*/4}{\sqrt{mp_i^*(1-p_i^*)/4}} \geq \frac{3}{10\sqrt{2}}\right] \quad (\delta = \tfrac{1}{10\sqrt{m}}, p_i^*(1-p_i^*) \geq \tfrac{2}{9}) \\
&\geq \Pr\left[g \geq \frac{3}{10\sqrt{2}}\right] - O(1/\sqrt{m/4}) \quad \text{(Lemma 14)} \\
&\geq 0.416 - O(1/\sqrt{m}) \quad \text{(CDF of Gaussian)} \\
&\geq \frac{1}{3}. \quad \text{(for sufficiently large } m\text{)}
\end{aligned}
$$

Thus, with probability at least 1/3, $\hat{\Delta}$ increases by at least $(mp_i^*/4 + m\delta/2) - (p_i^* + \delta) \cdot (m/4) = m\delta/4$ from time t_0 to t_1. This implies that either $\hat{\Delta}(t_0) \leq -m\delta/8$ or $\hat{\Delta}(t_1) \geq m\delta/8$. Note that both $\hat{\Delta}$ and $-\hat{\Delta}$ are lower bounds on $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p|$. So $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p|$ must reach $m\delta/8 = \sqrt{m}/80 = \frac{1}{1440}\sqrt{\frac{T}{k^\alpha \ln T}} = \theta$ at some point, and the epoch should have been terminated.

Proofs for P_2 and P_4. Again, we only present the proof for P_2 and the proof for P_4 is symmetric. Define $\hat{\Delta}(t) := \sum_{p \in P_2}[-\Delta_p(t)]$. When a probability $p \in P_2$ is predicted by the forecaster, $\hat{\Delta}$ is incremented by $p - b$, where $b \sim \mathrm{Ber}(p^*)$ and $p > p_i^* + \delta$. Therefore, the total increment in $\hat{\Delta}$ after the $m/4$ predictions is at least $(m/4)(p^* + \delta) - Z$, where $Z \sim B(m/4, p^*)$. Again, applying Lemma 14 gives $\Pr\left[Z \leq mp_i^*/4\right] \geq 1/3$ for sufficiently large T. When $Z \leq mp_i^*/4$, the increment would be at least $m\delta/4$, which further implies that $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p|$ should have reached $\theta = m\delta/8$ and the epoch should have terminated. □

Proof of Lemma 12. Let $\mathcal{I}_i = (l_i, r_i)$ be the interval associated with epoch i and $p_i^* = (l_i + r_i)/2$ be its middle point. Let t_0 be the last time step of epoch i. If epoch i is truthful and non-negligible, it holds that $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p(t_0)| \geq \theta$. We consider the following two cases depending on whether $\sum_{p \in P \cap \mathcal{I}_i} \Delta_p^+(t_0)$ or $\sum_{p \in P \cap \mathcal{I}_i} \Delta_p^-(t_0)$ is larger.

Case 1: $\sum_{p \in P \cap \mathcal{I}_i} \Delta_p^+(t_0) \geq \sum_{p \in P \cap \mathcal{I}_i} \Delta_p^-(t_0)$. As the two summations sum up to $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p(t_0)| \geq \theta$, we have $\sum_{p \in P \cap \mathcal{I}_i} \Delta_p^+(t_0) \geq \theta/2$. Let $P^+ := \{p \in P \cap \mathcal{I}_i : \Delta_p(t_0) \geq 0\}$. For each time step t, define the quantity $\hat{\Delta}(t)$ as

$$\hat{\Delta}(t) := \sum_{p \in P^+} \Delta_p(t).$$

Then, we have $\hat{\Delta}(t_0) = \sum_{p \in P^+} \Delta_p(t_0) = \sum_{p \in P \cap \mathcal{I}_i} \Delta_p^+(t_0) \geq \theta/2$.

For epoch i to be covered, when the scheme terminates at time T^{actual}, it should hold that $\sum_{p \in P \cap \mathcal{I}_i} |\Delta_p(T^{\text{actual}})| < \theta/4$. Since P^+ is a subset of $P \cap \mathcal{I}_i$ and $\Delta_p \leq |\Delta_p|$, it should also hold that $\hat{\Delta}(T^{\text{actual}}) = \sum_{p \in P^+} \Delta_p(T^{\text{actual}}) \leq \sum_{p \in P \cap \mathcal{I}_i} |\Delta_p(T^{\text{actual}})| < \theta/4$. Thus, $\hat{\Delta}$ needs to decrease by at least $\theta/4$ from time t_0 to T^{actual}.

Note that predictions with values outside P^+ does not affect $\hat{\Delta}$. Let m be the number of predictions that fall into P^+ strictly after epoch i (i.e., during time steps $t_0 + 1, t_0 + 2, \ldots, T^{\text{actual}}$). For each such prediction, suppose that it belongs to the i'-th epoch for some $i' > i$. Then, by Algorithm 1, the bit given by the scheme is drawn from $\mathrm{Ber}(p_{i'}^*)$, where $p_{i'}^*$ is the middle point of the interval $\mathcal{I}_{i'}$ associated with epoch i'. Recall that we assumed $\sum_{p \in P \cap \mathcal{I}_i} \Delta_p^+(t_0) \geq \sum_{p \in P \cap \mathcal{I}_i} \Delta_p^-(t_0)$ in this case, so in Algorithm 1, we let the forecaster place "+" into the cell in round i. For this sign to be preserved in the end, it must hold that the cell chosen in round i' has a larger number than the cell chosen in round i. By the choice of p_i^* and $p_{i'}^*$ in the scheme, we have $p_{i'}^* \geq p_i^* + \frac{1}{3k}$. Thus, the contribution of this prediction to $\hat{\Delta}$ is $b - p$, where $b \sim \mathrm{Ber}(p_{i'}^*)$ and $p \in \mathcal{I}_i = (p_i^* - \frac{1}{6k}, p_i^* + \frac{1}{6k})$. Thus, the increase in $\hat{\Delta}$ is bounded between -1 and 1 and has expectation

$$p_{i'}^* - p > \left(p_i^* + \frac{1}{3k}\right) - \left(p_i^* + \frac{1}{6k}\right) = \frac{1}{6k}.$$

Therefore, the total increment in $\hat{\Delta}$ over the m predictions inside P^+ is the sum of m random variables $X_1, X_2, \ldots, X_m$ satisfying that for each $j \in [m]$: (1) $X_j \in [-1, 1]$ almost surely; (2) $\mathbb{E}\left[X_j | X_1, X_2, \ldots, X_{j-1}\right] \geq \frac{1}{6k}$. Then, applying Lemma 13 gives

$$
\begin{aligned}
\Pr\left[\hat{\Delta}(T^{\text{actual}}) - \hat{\Delta}(t_0) \leq -\theta/4\right] &\leq \exp\left(-\frac{1}{2m}\left(\frac{m}{6k} + \frac{\theta}{4}\right)^2\right) \\
&\leq \exp\left(-\frac{\theta}{12k}\right).
\end{aligned}
$$

Finally, by a union bound over the $\leq T$ possible values of m, the probability that epoch i satisfies all the conditions in Lemma 12 is at most $T \cdot \exp\left(-\frac{\theta}{12k}\right)$. Plugging $k = T^{\frac{1}{\alpha+2\beta+2}}$ and $\theta = \frac{1}{1440}\sqrt{\frac{T}{k^\alpha \ln T}}$ into the bound shows that it is $o(1/T)$.

Case 2: $\sum_{p \in P \cap \mathcal{I}_i} \Delta_p^+(t_0) < \sum_{p \in P \cap \mathcal{I}_i} \Delta_p^-(t_0)$. This case is analogous to Case 1 and the proof is almost the same. Here we define $P^- := \{p \in P \cap \mathcal{I}_i : \Delta_p(t_0) < 0\}$ and $\hat{\Delta}(t) := \sum_{p \in P^-}[-\Delta_p(t)]$ instead. Then, for epoch i to be both non-negligible and covered, $\hat{\Delta}$ needs to be decreased from $\hat{\Delta}(t_0) \geq \theta/2$ to $\hat{\Delta}(T^{\text{actual}}) < \theta/4$. Suppose that exactly m predictions after epoch i fall into set P^-. Again, assuming that the sign placed in round i is preserved, we can show that the contribution of each such prediction to $\hat{\Delta}$ is always between -1 and 1 and has expectation at least $\frac{1}{6k}$. Therefore, applying Lemma 13 shows that the probability that $\hat{\Delta}_t$ decreases by at least $\theta/4$ after m such predictions is exponentially small, which completes the proof. □

REFERENCES

[1] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78, 1 (1950), 1–3.

[2] A Philip Dawid. 1982. The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* 77, 379 (1982), 605–610.

[3] Dean P Foster. 1999. A proof of calibration via Blackwell's approachability theorem. *Games and Economic Behavior* 29, 1-2 (1999), 73–78.

[4] Dean P Foster and Sergiu Hart. 2020. Forecast-Hedging and Calibration. http://www.ma.huji.ac.il/hart/papers/calib-int.pdf. Accessed: 2020-12-01.

[5] Dean P Foster, Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. 2011. Complexity-based approach to calibration with checking rules. In *Conference on Learning Theory (COLT)*. 293–314.

[6] Dean P Foster and Rakesh V Vohra. 1998. Asymptotic calibration. *Biometrika* 85, 2 (1998), 379–390.

[7] Drew Fudenberg and David K Levine. 1999. An easier way to calibrate. *Games and Economic Behavior* 29, 1-2 (1999), 131–137.

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning (ICML)*. 1321–1330.

[9] Sergiu Hart. 2020. Calibrated Forecasts: The Minimax Proof. http://www.ma.huji.ac.il/hart/papers/calib-minmax.pdf. Accessed: 2020-12-01.

[10] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning (ICML)*. 1939–1948.

[11] Christopher Jung, Changhwa Lee, Mallesh M Pai, Aaron Roth, and Rakesh Vohra. 2020. Moment Multicalibration for Uncertainty Estimation. *arXiv preprint arXiv:2008.08037* (2020).

[12] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Innovations in Theoretical Computer Science Conference (ITCS)*. 43:1–43:23.

[13] Volodymyr Kuleshov and Percy S Liang. 2015. Calibrated structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*. 3474–3482.

[14] Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3792–3803.

[15] Ehud Lehrer. 2001. Any inspection is manipulable. *Econometrica* 69, 5 (2001), 1333–1347.

[16] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems (NIPS)*. 5680–5689.

[17] Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J Candès. 2019. With Malice Towards None: Assessing Uncertainty via Equalized Coverage. *arXiv preprint arXiv:1908.05428* (2019).

[18] Alvaro Sandroni, Rann Smorodinsky, and Rakesh V Vohra. 2003. Calibration with many checking rules. *Mathematics of Operations Research* 28, 1 (2003), 141–153.

[19] Eliran Shabat, Lee Cohen, and Yishay Mansour. 2020. Sample Complexity of Uniform Convergence for Multicalibration. *arXiv preprint arXiv:2005.01757* (2020).

[20] Vladimir Vovk. 2007. Non-asymptotic calibration and resolution. *Theoretical Computer Science* 387, 1 (2007), 77–89.

[21] Shengjia Zhao, Tengyu Ma, and Stefano Ermon. 2020. Individual Calibration with Randomized Forecasting. In *International Conference on Machine Learning (ICML)*. 8366–8376.