UCLA
UCLA Electronic Theses and Dissertations

Title

On the Distance from Calibration in Sequential Prediction

Permalink

https://escholarship.org/uc/item/5gx9g01k

Author

ZHENG, LETIAN

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

On the Distance from Calibration in Sequential Prediction

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Computer Science

by

Letian Zheng

2024

© Copyright by

Letian Zheng

2024

ABSTRACT OF THE THESIS

On the Distance from Calibration

in Sequential Prediction

by

Letian Zheng

Master of Science in Computer Science

University of California, Los Angeles, 2024

Professor Quanquan Gu, Chair

We study a sequential binary prediction setting where the forecaster is evaluated in terms of the *calibration distance*, which is defined as the L_1 distance between the predicted values and the set of predictions that are perfectly calibrated in hindsight. This is analogous to a calibration measure recently proposed by Błasiok, Gopalan, Hu and Nakkiran (STOC 2023) for the offline setting. The calibration distance is a natural and intuitive measure of deviation from perfect calibration, and satisfies a Lipschitz continuity property which does not hold for many popular calibration measures, such as the L_1 calibration error and its variants.

We prove that there is a forecasting algorithm that achieves an $O(\sqrt{T})$ calibration distance in expectation on an adversarially chosen sequence of T binary outcomes. At the core of this upper bound is a structural result showing that the calibration distance is accurately approximated by the *lower calibration distance*, which is a continuous relaxation of the former. We then show

that an $O(\sqrt{T})$ lower calibration distance can be achieved via a simple minimax argument and a reduction to online learning on a Lipschitz class.

On the lower bound side, an $\Omega(T^{1/3})$ calibration distance is shown to be unavoidable, even when the adversary outputs a sequence of independent random bits, and has an additional ability to *early stop* (i.e., to stop producing random bits and output the same bit in the remaining steps). Interestingly, without this early stopping, the forecaster can achieve a much smaller calibration distance of polylog(T).

The thesis of Letian Zheng is approved.

Adnan Youssef Darwiche

Raghu Meka

Quanquan Gu, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

Acknowledgement

This paper is directed by Dr. Mingda Qiao, who raises the core problems based on the paper by Błasiok, Gopalan, Hu and Nakkiran (STOC 2023) and his previous research in calibration error. I sincerely appreciate his patient elaboration of any problems or theories that confused me and unreserved support throughout the whole project, which led me to the world of scientific research. His enthusiasm inspired me to devote myself to the research and fully contribute my sparks of inspiration in problem solving to our theorems and algorithms. Besides, he demonstrated an exceptional level of academic rigor in the project as he perfected the proofs in the paper, guiding me to transform our concrete examples and strategies into abstract validation of theorems and definitions of algorithms.

Dr. Mingda Qiao raised the problems, advanced our theorems and algorithms, composed the main textual sections, including abstract and introduction, and wrote or finalized most of the proofs of theorems.

On the Distance from Calibration in Sequential Prediction

Mingda Qiao* Letian Zheng†

Abstract

We study a sequential binary prediction setting where the forecaster is evaluated in terms of the *calibration distance*, which is defined as the L_1 distance between the predicted values and the set of predictions that are perfectly calibrated in hindsight. This is analogous to a calibration measure recently proposed by Błasiok, Gopalan, Hu and Nakkiran (STOC 2023) for the offline setting. The calibration distance is a natural and intuitive measure of deviation from perfect calibration, and satisfies a Lipschitz continuity property which does not hold for many popular calibration measures, such as the L_1 calibration error and its variants.

We prove that there is a forecasting algorithm that achieves an $O(\sqrt{T})$ calibration distance in expectation on an adversarially chosen sequence of T binary outcomes. At the core of this upper bound is a structural result showing that the calibration distance is accurately approximated by the *lower calibration distance*, which is a continuous relaxation of the former. We then show that an $O(\sqrt{T})$ lower calibration distance can be achieved via a simple minimax argument and a reduction to online learning on a Lipschitz class.

On the lower bound side, an $\Omega(T^{1/3})$ calibration distance is shown to be unavoidable, even when the adversary outputs a sequence of independent random bits, and has an additional ability to *early stop* (i.e., to stop producing random bits and output the same bit in the remaining steps). Interestingly, without this early stopping, the forecaster can achieve a much smaller calibration distance of polylog(T).

1 Introduction

We revisit the sequential binary prediction setup of Foster and Vohra [FV98], in which a forecaster makes probabilistic predictions on a sequence of T adversarially chosen binary outcomes. At each step $t \in [T]$, the adversary picks a bit $x_t \in \{0, 1\}$ and, simultaneously, the forecaster makes a prediction $p_t \in [0, 1]$ on the "probability" of $x_t = 1$. These values are then revealed to both players, and may factor into their subsequent actions.

The forecaster is evaluated in terms of the *calibration* criterion, which is a natural and intuitive condition for the predictions to be interpretable as probabilities. The predictions are called *perfectly calibrated* if, among the steps on which each $\alpha \in [0, 1]$ is predicted, exactly an α fraction of the bits are ones. Formally, it must hold for every $\alpha \in [0, 1]$ that $\sum_{t=1}^{T}(x_t - p_t) \cdot \mathbb{1}[p_t = \alpha] = 0$.

Quantitatively, the L_1 calibration error, also known as the Expected Calibration Error (ECE),

*University of California, Berkeley. Email: mingda.qiao@berkeley.edu. Part of this work was done while the author was a graduate student at Stanford University.

†University of California, Los Angeles. Email: letianzh@g.ucla.edu.

is defined as the total violation of calibration over all $\alpha \in [0,1]$[1]:

$$\mathsf{ECE}(x,p) \coloneqq \sum_{\alpha \in [0,1]} \left| \sum_{t=1}^{T} (x_t - p_t) \cdot \mathbb{1}\left[p_t = \alpha\right] \right|. \tag{1}$$

While this definition seems natural, the ECE can be *ill-behaved* because it is discontinuous with respect to the predictions. For instance, when x contains the same number of 0s and 1s, predicting $p_t = 1/2$ at every $t \in [T]$ achieves a zero ECE. However, if we replace each p_t with $1/2 + \epsilon_t$, where $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$ are arbitrarily small, non-zero, and distinct perturbations, the ECE suddenly jumps to $\Omega(T)$, as the T steps count as T different "bins" in Equation (1). As noted by [BGHN23], while this issue can be alleviated by binning the prediction values, the binning may introduce discontinuity at the boundary of each sub-interval, and there is no consensus on how the binning should be chosen in general.

In this work, we study a variant of this fundamental sequential calibration setup, in which the distance from calibration is defined as the minimum L_1 distance between the predicted values and the closest predictions that are perfectly calibrated with respect to the outcomes. This definition is analogous to a calibration measure recently proposed and studied by [BGHN23] for the offline setting. Formally, with respect to outcomes $x_1, x_2, \ldots, x_T$, the *calibration distance* of predictions $p_1, p_2, \ldots, p_T$ is defined as

$$\mathsf{CalDist}(x,p) \coloneqq \min_{q \in \mathcal{C}(x)} \|p - q\|_1,$$

where $\mathcal{C}(x) \coloneqq \left\{ p \in [0,1]^T : \forall \alpha \in [0,1],\ \sum_{t=1}^{T} (x_t - p_t) \cdot \mathbb{1}\left[p_t = \alpha\right] = 0 \right\}$ is the set of predictions that are perfectly calibrated for x.[2]

Equivalently, the calibration distance measures the minimum amount of modification that the forecaster has to make to its predictions, in order for them to be calibrated in hindsight. It follows immediately from the definition that the calibration distance is robust to small perturbations in the predictions, and thus avoids the discontinuity issue of the ECE. In Appendix A, we prove that the calibration distance is always upper bounded by the ECE, so the calibration distance can also be viewed as a relaxation of the ECE.

In this work, we address the following two questions regarding this new calibration measure.

Question 1. *Can we efficiently compute (or at least approximate) the calibration distance on given outcomes and predictions?*

Question 2. *What is the optimal calibration distance that the forecaster can guarantee against T adversarially chosen outcomes?*

1.1 Overview of Our Results

Efficient approximation via a structural result. We start by giving a positive answer to Question 1, up to a small additive approximation error.

[1] The summand is non-zero only if $\alpha \in \{p_1, p_2, \ldots, p_T\}$, so the summation is finite and well-defined. In the rest of the paper, we frequently abuse the notation $\sum_{\alpha \in [l,r]} x_\alpha$ if the x_α is non-zero on finitely many values of $\alpha \in [l,r]$.

[2] Note that every $p \in \mathcal{C}(x)$ corresponds to a partition of $[T]$, so $\mathcal{C}(x)$ is a finite set of size $T^{O(T)}$. Therefore, the minimum in the definition of $\mathsf{CalDist}$ can be achieved.

Theorem 1. *There is an algorithm that, given $x \in \{0,1\}^T$ and $p \in [0,1]^T$, outputs an estimate of* $\mathsf{CalDist}(x,p)$ *up to an* $O(\sqrt{T})$ *additive error in* $\mathrm{poly}(T)$ *time.*

We prove Theorem 1 by relating $\mathsf{CalDist}(x,p)$ to the *lower calibration distance*, denoted by $\mathsf{LowerCalDist}(x,p)$, which we formally define in Section 2. Roughly speaking, the definition of $\mathsf{LowerCalDist}(x,p)$ allows us to compare $p_1, p_2, \ldots, p_T$ to *randomized predictions* $q_1, q_2, \ldots, q_T$. For the offline setting, [BGHN23] introduced an analogous notion, and gave a $\mathrm{poly}(T, 1/\epsilon)$ time algorithm that approximates the lower calibration distance up to any additive error of $\epsilon > 0$. Therefore, Theorem 1 would immediately follow if we could show that $\mathsf{LowerCalDist}(x,p)$ is a good approximation of $\mathsf{CalDist}(x,p)$ for any x and p.

A result of [BGHN23] implies that, after normalizing by a $1/T$ factor, these two measures are indeed polynomially related:

$$\frac{\mathsf{LowerCalDist}(x,p)}{T} \le \frac{\mathsf{CalDist}(x,p)}{T} \le 4\sqrt{\frac{\mathsf{LowerCalDist}(x,p)}{T}}.$$

On the other hand, this quadratic gap is unavoidable in general: We show in Proposition 4 that even for $T = 4$, there exist x and p such that $\mathsf{CalDist}(x,p) = \Omega(\epsilon)$ but $\mathsf{LowerCalDist}(x,p) = O(\epsilon^2)$ for sufficiently small $\epsilon > 0$. Taking $\epsilon \to 0^+$ shows that $\mathsf{LowerCalDist}(x,p)$ is not a good *multiplicative* approximation of $\mathsf{CalDist}(x,p)$.

Fortunately, in the example above, the *additive* gap between the two calibration measures is small. Our key technical result below states that this is true in general: $\mathsf{CalDist}(x,p)$ is always upper bounded by $\mathsf{LowerCalDist}(x,p) + O(\sqrt{T})$. Thus, Theorem 1 indeed follows from the algorithm of [BGHN23] for approximating $\mathsf{LowerCalDist}(x,p)$. Furthermore, when p is *sparse* in the sense that it contains only a few different entries, we improve the additive gap from $\sqrt{T}$ to the sparsity level, at the cost of an extra constant factor.

Theorem 2. *For any $x \in \{0,1\}^T$ and $p \in [0,1]^T$, we have:*

- $\mathsf{CalDist}(x,p) \le \mathsf{LowerCalDist}(x,p) + O(\sqrt{T})$.
- $\mathsf{CalDist}(x,p) \le O(1) \cdot \mathsf{LowerCalDist}(x,p) + O(m)$, *where* $m = |\{p_1, p_2, \ldots, p_T\}|$.

The $O(\cdot)$ notations above hide universal constant factors that are independent of T, x, and p.

Upper bound via a minimax argument. Our next result addresses Question 2 from the upper bound side.

Theorem 3. *There is forecasting algorithm that, against any adversary, achieves an $O(\sqrt{T})$ calibration distance in expectation.*

In light of Theorem 2, it suffices to give a forecaster with an expected lower calibration error of $O(\sqrt{T})$. [BGHN23] showed that the lower calibration distance and the *smooth calibration error* (which we define in Section 2) differ by a constant factor. Therefore, at the core of our proof of Theorem 2 is an $O(\sqrt{T})$ upper bound on the smooth calibration error, which is proved via a minimax argument similar to the proof of Hart for upper bounding the ECE [FV98, Har22]. We note that Kakade and Foster [KF08] gave an algorithm with a sub-linear smooth calibration error, though directly following their proof gives a looser upper bound of $O(T^{2/3})$.

Impossibility of impossibility results from random bits. It might appear "obvious" that the $O(\sqrt{T})$ bound in Theorem 3 is tight: Suppose that the adversary plays a sequence of T independent random bits. Intuitively, the forecaster's best strategy is to predict $p_t = 1/2$ at every step t. Then, the calibration distance can be shown to be $\Omega(\sqrt{T})$ in expectation.[3]

Surprisingly, this argument turns out to be incorrect—in fact, "exponentially" incorrect![4]

Proposition 1. *When the adversary promises to play T independent random bits, there is a forecasting algorithm that achieves an $O(\log^{3/2} T)$ calibration distance in expectation.*

Our proof of Proposition 1 has two steps: First, we give a strategy that achieves a small smooth calibration error. This is done by first predicting $1/2$ and then, based on the realization of the random bits, predicts a slightly biased value in the hope of de-biasing the previous mistakes. To translate this to an upper bound on the calibration distance, the first bound in Theorem 2 is insufficient, since the $\sqrt{T}$ gap would dominate the polylog(T) error. Fortunately, in the first step, we always predict at most $O(\log T)$ different values, so the second bound in Theorem 2 can be applied instead.

Lower bounds from random bits. Despite the surprising fact above, we still manage to prove a poly(T) lower bound.

Theorem 4. *There is a strategy for the adversary such that any forecasting algorithm must incur an $\Omega(T^{1/3})$ calibration distance in expectation.*

Theorem 4 is proved by providing a minimal additional ability to the adversary that produces random bits. The new adversary will generate random bits until the calibration distance hits $\Omega(T^{1/3})$ at some point, and then either keep playing zeros or keep playing ones, depending on which bit could ensure that the calibration distance is still large in the end.

1.2 Related Work

Calibration is a natural criterion for evaluating probabilistic forecasts. The idea of calibration can be at least traced back to Brier [Bri50]. A formal definition of calibration appeared in the work of Dawid [Daw82, Daw85]. There are huge bodies of recent work on the calibration of neural networks [GPSW17] and the use of calibration (and its extension such as multi-calibration) as a measure of algorithmic fairness [KMR17, HJKRR18]. In the following, we focus the discussion mainly on calibration in sequential setups, which is the closest to this paper.

Distance from calibration. In the context of offline probabilistic prediction, [BGHN23] noted that while "the notion of perfect calibration is well-understood", "there is no consensus" on how the distance from perfect calibration should be quantified. They proposed to use the following as the ground truth for the distance of a predictor from calibration: the minimum ℓ_1 distance between the predictor and any predictor that is perfectly calibrated with respect to the underlying distribution. The authors then examined various calibration measures, and identified which of them are *consistent* in the sense that of being polynomially related to this ground truth.

[3] This follows from $\mathsf{CalDist}(x,p) \geq \mathsf{LowerCalDist}(x,p) \geq \frac{1}{2}\mathsf{smCE}(x,p) \geq \frac{1}{2}\left|\sum_{t=1}^{T}(x_t - p_t)\right|$ and that the last term is $\Omega(\sqrt{T})$ in expectation; see Section 2 for justification of the first three steps.

[4] Nevertheless, this argument indeed shows that the proof strategy of Theorem 3 at best gives an $O(\sqrt{T})$ bound.

In the sequential setup, however, even the notion of "perfect calibration" might be at odds with what intuitively count as the "right" predictions. It is easy to construct examples in which the adversary generates the outcomes randomly from a known distribution, yet the only way to achieve perfect calibration is through "lying" on some predictions, i.e., predicting a value that is far from the true (conditional) probability of the next outcome (see, e.g., [QV21, Example 2] for a concrete instance).

On a technical level, the notion of consistency in the study of [BGHN23] might be loose by a quadratic factor (which is also shown to be unavoidable in their formulation). In contrast, the goal of this work is to pin down the optimal rate of the calibration distance, so this quadratic gap poses a challenge. Roughly speaking, the quadratic gap that is unavoidable in the setup of [BGHN23] is due to the uncertainty in the granularity of the marginal distribution. This gap is partially avoided in the sequential setup, since the "marginal" is always the uniform distribution over $[T]$.

Sequential calibration and variants. Foster and Vohra [FV98] gave the first algorithm that achieves a vanishing (squared L_2) calibration error as $T \to +\infty$ on a sequence of T adversarially generated outcomes. Alternative proofs were subsequently given by [FL99, Fos99]. In terms of the error rates, the optimal ECE (defined in Equation (1)) is known to be between $O(T^{2/3})$ (implict in [FV98]; see [Har22] for a formal exposition) and $\Omega(T^{0.528})$ [QV21].

[FRST11] studied a strengthened notion of calibration that requires the predictions to be calibrated even when restricted to certain subsets of the time horizon (also called "checking rules"). They derived convergence bounds that depend on different complexity measures of the family of checking rules.

Relaxed versions of the ECE, including weak calibration [KF08], smooth calibration [FH18] and continuous calibration [FH21], have also been studied. These alternative calibration notions also resolve the discontinuity issue of ECE, and were shown to be achievable by deterministic forecasting algorithms. In contrast, any algorithm with a sub-linear ECE must be randomized. [GR22] studied a "power of two choice" variant in which the forecaster is allowed to predict two different (yet nearby) values at each step, and use the one closer to the outcome after the outcome is revealed.

Calibration-accuracy trade-off. Another variant of the problem is when the forecaster is given a hint or expert advice before predicting at each step. The goal is to re-calibrate the expert's predictions without increasing the cumulative loss in the predictions. [KE17, OKS23] gave trade-offs between the ECE incurred by the forecaster and the *regret*, defined as the excess loss compared to always following the hints.

Online multi-calibration. Multi-calibration is a stronger notion of calibration proposed by [HJKRR18] in the context of fairness of machine learning models in offline setups. This notion requires the predictions to be calibrated on a family of pre-specified subsets of the feature space as well. A recent line of work [GJN+22, LNPR22, BGJ+22, GJRR24] gave algorithms that achieve approximate multi-calibration in the online setup, in which the features and labels are sequentially and adversarially chosen.

Calibrated predictions for decision-making. Recent work of [KLST23] and [NRRX23] studied models in which the predictions are used for downstream decision-making. [KLST23] defined "U-Calibration", which is shown to be equivalent to the sub-linear regret guarantee for *all* decision

makers. [NRRX23] gave algorithms that are calibrated even when evaluated in conjunction with the decisions, which might, in turn, depend on the predictions.

1.3 Organization of the Paper

In Section 2, we formally introduce two calibration measures in the literature, the *lower calibration distance* and the *smooth calibration error*, both of which are closely related to the calibration distance, and will play key roles in our proofs. In Section 3, we sketch the proofs of our main results. We suggest that the readers read this section before delving into the formal proofs, as most of the proofs are based on simple ideas and intuition that need slightly heavier notations to be formalized. We discuss the suitability of the calibration distance as a calibration measure, and highlight a few open problems in Section 4. The formal proofs are given in Sections 5 through 8.

2 Preliminaries

The binary outcomes and the predictions are denoted by $x_1, x_2, \ldots, x_T$ and $p_1, p_2, \ldots, p_T$. We use the shorthands $x_{l:r}$ and $p_{l:r}$ for subsequences $x_l, x_{l+1}, \ldots, x_r$ and $p_l, p_{l+1}, \ldots, p_r$. For $\alpha \in [0,1]$ and $t \in \{0, 1, 2, \ldots, T\}$, $\Delta_\alpha(t) \coloneqq \sum_{t'=1}^{t} (x_{t'} - p_{t'}) \cdot \mathbb{1}[p_{t'} = \alpha]$ is the total bias that the forecaster incurs on prediction value α up to time t. We drop the argument t when it is clear from the context.

Lower calibration distance. Below is a formal definition of the lower calibration distance, which is equivalent to the "lower distance from calibration" defined by [BGHN23] up to a normalization factor of T.

Definition 1 (Lower Calibration Distance)**.** *The lower calibration distance of predictions $p \in [0,1]^T$ with respect to outcomes $x \in \{0,1\}^T$ is*

$$\mathsf{LowerCalDist}(x, p) \coloneqq \inf_{\mathcal{D} \in \underline{\mathcal{C}}(x)} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \mathcal{D}_t} [|p_t - q_t|],$$

where $\underline{\mathcal{C}}(x)$ is the family of T-tuples of distributions $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T)$ such that: (1) The support of each $\mathcal{D}_t$ is finite and contained in $[0,1]$; (2) $\mathcal{D}$ is perfectly calibrated with respect to x in the sense that $\sum_{t=1}^{T} (x_t - \alpha) \cdot \mathcal{D}_t(\alpha) = 0$ holds for every $\alpha \in [0,1]$.

We will use the shorthand $\|p - \mathcal{D}\|_1 \coloneqq \sum_{t=1}^{T} \mathbb{E}_{q_t \sim \mathcal{D}_t} [|p_t - q_t|]$ for $p \in [0,1]^T$ and distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ over $[0,1]$. The definition above can then be simplified to $\inf_{\mathcal{D} \in \underline{\mathcal{C}}(x)} \|p - \mathcal{D}\|_1$.

Remark 2. *Definition 1 becomes more natural when viewed through the lens of optimal transport. Imagine that, for each $t \in [T]$, there is one unit of bit x_t located at point p_t. Then, each distribution $\mathcal{D}_t$ specifies a way of splitting and transporting the mass to (finitely many) different locations on $[0,1]$. The distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ are in the family $\underline{\mathcal{C}}(x)$ if and only if after all the transportations are done, at every location $\alpha \in [0,1]$, the fraction of ones is exactly α. (Indeed, the constraint $\sum_{t=1}^{T} (x_t - \alpha) \cdot \mathcal{D}_t(\alpha) = 0$ is equivalent to $\frac{\sum_{t=1}^{T} x_t \cdot \mathcal{D}_t(\alpha)}{\sum_{t=1}^{T} \mathcal{D}_t(\alpha)} = \alpha$.) The lower calibration distance is exactly the minimum cost of the transportation subject to the calibration constraint on the resulting configuration, when the cost of moving one unit of mass from p_t to q_t is given by $|p_t - q_t|$.*

In comparison, the definition of the calibration distance introduces an additional constraint: each unit of mass cannot be transported to multiple locations, i.e., each $\mathcal{D}_t$ must be a degenerate distribution. This immediately gives the inequality $\mathsf{LowerCalDist}(x, p) \le \mathsf{CalDist}(x, p)$.

Smooth calibration error. Another related calibration measure is the *smooth calibration error* proposed by [KF08]:

$$\mathsf{smCE}(x,p) \coloneqq \sup_{f\in\mathcal{F}} \sum_{t=1}^{T} f(p_t)(x_t - p_t) = \sup_{f\in\mathcal{F}} \sum_{\alpha\in[0,1]} f(\alpha)\cdot\Delta_\alpha(T),$$

where $\mathcal{F}$ is the family of 1-Lipschitz functions from $[0,1]$ to $[-1,1]$.

It was shown by [BGHN23] that the smooth calibration error and the lower calibration distance are at most a constant factor away.

Lemma 3 (Theorem 7.3 of [BGHN23])**.** *For any $x \in \{0,1\}^T$ and $p \in [0,1]^T$,*

$$\frac{1}{2}\mathsf{smCE}(x,p) \le \mathsf{LowerCalDist}(x,p) \le 2\mathsf{smCE}(x,p).$$

Lipschitz continuity. Unlike the ECE, $\mathsf{CalDist}(x,p)$, $\mathsf{LowerCalDist}(x,p)$, and $\mathsf{smCE}(x,p)$ are all Lipschitz in the predictions p. Formally, we prove in Appendix A that, for fixed $x \in \{0,1\}^T$ and with respect to the 1-norm, $\mathsf{CalDist}(x,p)$ and $\mathsf{LowerCalDist}(x,p)$ are 1-Lipschitz, while $\mathsf{smCE}(x,p)$ is 2-Lipschitz.

3 Proof Overview

We sketch the proofs of our results in this section. The proof of the approximation guarantees (Theorem 2) is the most involved and consists of several technical ingredients, for which we give an overview in Section 3.1. Given this approximation guarantee, the $O(\sqrt{T})$ upper bound (Theorem 3) follows from a minimax argument and a reduction to online learning, outlined in Section 3.2.

Both the polylog(T) upper bound for random bits (Proposition 1) and the $\Omega(T^{1/3})$ lower bound based on random bits and early stopping (Theorem 4) are based on abstracting the setup as a "controlled random walk" game, which we define in Section 3.3. We will discuss how to solve the game with a polylog(T) cost, and why that translates into an upper bound on the calibration distance. Finally, in Section 3.4, we explain why a connection in the other direction also holds, and how the $\Omega(T^{1/3})$ lower bound follows.

3.1 Approximation Guarantees

To show that $\mathsf{LowerCalDist}(x,p)$ is a good approximation of $\mathsf{CalDist}(x,p)$, we first pick $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T) \in \underline{\mathcal{C}}(x)$ as a "witness" of $\mathsf{LowerCalDist}(x,p)$, i.e., $\mathsf{LowerCalDist}(x,p) = \|p - \mathcal{D}\|_1$.[5] Then, we round the distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ to deterministic values $q_1, \ldots, q_T \in [0,1]$ such that $q \in \mathcal{C}(x)$ and $\|p-q\|_1 \le \alpha\|p-\mathcal{D}\|_1 + \beta$. The desired approximation guarantee would then follow from

$$\mathsf{CalDist}(x,p) \le \|p-q\|_1 \le \alpha\|p-\mathcal{D}\|_1 + \beta = \alpha\mathsf{LowerCalDist}(x,p) + \beta.$$

This rounding is done in two steps. First, we transform $\mathcal{D}$ into another sequence $\mathcal{D}' = (\mathcal{D}'_1, \ldots, \mathcal{D}'_T)$ of T distributions, such that: (1) $D' \in \underline{\mathcal{C}}(x)$, i.e., $\mathcal{D}'$ is still perfectly calibrated;

[5] Technically, we can only find $\mathcal{D}$ that achieves $\|p-\mathcal{D}\|_1 \le \mathsf{LowerCalDist}(x,p) + \epsilon$ for some $\epsilon > 0$. Since ϵ can be made arbitrarily small, the rest of the argument would not be affected.

(2) There is a small finite set $S \subset [0,1]$ that contains the support of every $\mathcal{D}'_t$; (3) $\|p - \mathcal{D}'\|_1$ can be upper bounded in terms of $\|p - \mathcal{D}\|_1$.

Concretely, Lemma 6 gives such a transformation that guarantees

$$\|p - \mathcal{D}'\|_1 \le \|p - \mathcal{D}\|_1 + O(\sqrt{T}) \quad \text{and} \quad |S| = O(\sqrt{T}).$$

When p contains at most m different entries, Lemma 7 shows that we can alternatively achieve

$$\|p - \mathcal{D}'\|_1 \le O(1) \cdot \|p - \mathcal{D}\|_1 \quad \text{and} \quad |S| = O(m).$$

The final ingredient is a method of rounding distributions $\mathcal{D}'_1, \dots, \mathcal{D}'_T$ over set S to deterministic values $(q_1, \dots, q_T) \in \mathcal{C}(x)$. In Lemma 5, we give a rounding scheme with the guarantee $\|p - q\|_1 \le \|p - \mathcal{D}'\|_1 + O(|S|)$. Clearly, Lemmas 5 through 7 together prove Theorem 2.

General reduction to small support size. For the general case, we prove Lemma 6 using a simple binning strategy. Recall from Remark 2 that $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_T)$ specifies a way of transporting T units of mass (labeled with either 0 or 1) over the interval $[0,1]$. For each t, a unit amount of bit x_t is originally located at p_t, and gets transported according to distribution $\mathcal{D}_t$. The transportation incurs a total cost of $\|p - \mathcal{D}\|_1$, and the condition $\mathcal{D} \in \underline{\mathcal{C}}(x)$ requires that, in the resulting configuration, the fraction of ones at each $\alpha \in [0,1]$ is exactly α.

A priori, the bits might be transported to many different destinations. We partition the interval $[0,1]$ into $\sqrt{T}$ intervals with equal lengths. For the i-th interval $\left[\frac{i-1}{\sqrt{T}}, \frac{i}{\sqrt{T}}\right]$, we examine the mass being transported to all the locations within the interval. We *consolidate* these transportations by redirecting them to a single destination, which is chosen such that the calibration constraint is still satisfied. Clearly, there will be at most $\sqrt{T}$ different destinations after the consolidation for all the intervals. Since $\mathcal{D}$ is perfectly calibrated, it is easy to show that the new destination falls into the same interval as the original destinations do, and is thus at a distance $\le 1/\sqrt{T}$. The total increase in the transportation cost will be bounded by $T \cdot (1/\sqrt{T}) = \sqrt{T}$ as desired.

Reduction to small support size under sparsity. In the setup of Lemma 7, each p_t is one of the m values $s_1 < s_2 < \cdots < s_m$. In light of the proof strategy for the general case, it is tempting to try the following: Divide $[0,1]$ into $m+1$ intervals by splitting at each s_i. For each interval $[s_i, s_{i+1}]$, again, we consolidate all the transportations into the interval by redirecting them to a single location. Unfortunately, this does not work, since a typical interval $[s_i, s_{i+1}]$ has length $\Omega(1/m)$ and the redirection could incur an $\Omega(T/m)$ cost, which is too large.

In our proof, we still examine the mass being transported into the interval $[s_i, s_{i+1}]$ according to $\mathcal{D}$. Since each unit of mass originates at some $p_t \in \{s_1, s_2, \dots, s_m\}$, the origin must be in $[0, s_i] \cup [s_{i+1}, 1]$. An important simplifying observation is that we may assume that all the bits originate from either s_i or s_{i+1}. This is because the transportation of mass from some origin $p_t \in [0, s_i] \cup [s_{i+1}, 1]$ to a destination inside $[s_i, s_{i+1}]$ can be viewed as a two-phase process: first, transport the mass from p_t to one of the endpoints (s_i if $p_t \in [0, s_i]$ and s_{i+1} if $p_t \in [s_{i+1}, 1]$); then, transport it from the endpoint to the actual destination. We will keep the first phase of each transportation unchanged, and focus on consolidating the second phases, in which the origins are either s_i or s_{i+1}. We will ensure that, after the consolidation, there are $O(1)$ different destinations for each interval $[s_i, s_{i+1}]$, while the total cost of the second phases increases by at most an $O(1)$ factor. To find such a consolidation strategy, we exploit the connection between the lower calibration distance and smooth calibration error (Lemma 3), and perform a quite involved case analysis.

Rounding of distributions supported over a small set. Finally, we sketch the proof of our rounding lemma (Lemma 5). The starting point is a sequence of T distributions $\mathcal{D}_1, \dots, \mathcal{D}_T$ over a common set S of a small size. Let $s_1 < s_2 < \cdots < s_{|S|}$ be the elements of S. Suppose that for some $t_1 \neq t_2$, we have $x_{t_1} = x_{t_2}$, $p_{t_1} < p_{t_2}$. Meanwhile, $\mathcal{D}_{t_1}(s_i)$ and $\mathcal{D}_{t_2}(s_j)$ are both positive for some $i > j$. Intuitively, this means that $\mathcal{D}$ is inefficient—if we redirect an ϵ probability mass of $\mathcal{D}_{t_1}$ from s_i to s_j, and the same amount in $\mathcal{D}_{t_2}$ from s_j to s_i, we would end up with the same outcome without increasing the cost. In general, we should expect $\mathcal{D}$ to satisfy the following monotonicity property, or we can tweak it without increasing $\|p - \mathcal{D}\|_1$: for any t_1, t_2 such that $x_{t_1} = x_{t_2}$ and $p_{t_1} < p_{t_2}$, every element in the support of $\mathcal{D}_{t_1}$ is less than or equal to every element in the support of $\mathcal{D}_{t_2}$. In fact, this is a simple characterization of the optimal transport on a line.

Once we enforce this monotonicity, the rounding is easy—simply because there will not be much for us to round! Indeed, whenever $\mathcal{D}_t$ has a support of size at least 2 (say, $\{s_i, s_{i+1}\}$), step t must be, among all t' such that $x_{t'} = x_t$ and $\mathcal{D}_{t'}(s_{i+1}) > 0$, the one with the smallest value of $p_{t'}$. This shows that $\mathcal{D}_t$ is degenerate, except for $O(|S|)$ different choices of t. For each non-degenerate distribution $\mathcal{D}_t$, we naïvely pick $q_t = x_t$, so that calibration is satisfied. We also need to change the non-degenerate distributions to maintain calibration. It turns out that this rounding incurs an additional cost of $O(|S|)$, as desired.

3.2 Calibration Distance Upper Bound

Theorem 2 and Lemma 3 together give

$$\mathsf{CalDist}(x, p) \le \mathsf{LowerCalDist}(x, p) + O(\sqrt{T}) \le 2\mathsf{smCE}(x, p) + O(\sqrt{T}).$$

Therefore, to prove Theorem 3, it suffices to achieve an $O(\sqrt{T})$ smooth calibration error.

Suppose that the forecaster and the adversary are playing a zero-sum game, with the objective being the smooth calibration error. By the minimax theorem[6], we may assume that the adversary's (mixed) strategy is known. Then, at each step $t \in [T]$, we may compute the probability for the adversary to play $x_t = 1$ conditioning on $x_{1:(t-1)}$ and $p_{1:(t-1)}$. A natural strategy is then to choose p_t as this conditional probability. To analyze the smooth calibration error incurred by this strategy, we frame this game as an instance of online learning on the class $\mathcal{F}$ of Lipschitz functions from $[0, 1]$ to $[-1, 1]$, and apply a regret bound in the online learning literature. The $O(\sqrt{T})$ bound follows from the fact that $\mathcal{F}$ has an $O(\sqrt{T})$ *sequential Rademacher complexity*, which is an analogue of the usual Rademacher complexity for the sequential setup.

3.3 Improved Forecasters for Random Bits

When the adversary commits to producing T random bits, the minimization of the smooth calibration error is, informally, captured by the following control problem:

> **Controlled Random Walk:** The player starts at location $X_0 = 0$. At each step $t \in [T]$, the player first moves by $\epsilon_t \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ (which may depend on X_{t-1}). Then, the nature perturbs the player's location by $\delta_t \in \left\{\pm\frac{1}{2}\right\}$ chosen uniformly at random. In other words, the player is located at $X_t = X_{t-1} + \epsilon_t + \delta_t$ after step t. The *cost* of the player is defined as $|X_T| + \sum_{t=1}^{T} \epsilon_t^2$. What is the lowest possible expected cost?

[6]Technically, we need to restrict the predictions to a finite set to apply the minimax theorem. This is handled by rounding the predictions to a $1/T$-net of $[0, 1]$ and applying Lipschitz continuity in the formal proof.

To see how the game defined above is related to the calibration setup, recall that the smooth calibration error can be written as $\sup_{f\in\mathcal{F}} \sum_{\alpha\in[0,1]} f(\alpha)\cdot\Delta_\alpha$, where $\Delta_\alpha = \sum_{t=1}^T (x_t - p_t)\cdot \mathbb{1}[p_t = \alpha]$ is the total bias associated with prediction value α, and $\mathcal{F}$ is the family of 1-Lipschitz functions from $[0,1]$ to $[-1,1]$. For any $f\in\mathcal{F}$, we have

$$\begin{aligned}\sum_{\alpha\in[0,1]} f(\alpha)\cdot\Delta_\alpha &= f(1/2)\cdot\sum_{\alpha\in[0,1]}\Delta_\alpha + \sum_{\alpha\in[0,1]}[f(\alpha)-f(1/2)]\cdot\Delta_\alpha \\ &\le \left|\sum_{t=1}^T (x_t - p_t)\right| + \sum_{\alpha\in[0,1]} |\alpha - 1/2|\cdot|\Delta_\alpha|.\end{aligned}$$

If we write $p_t = 1/2 - \epsilon_t$ and $x_t = 1/2 + \delta_t$, the first term above reduces to $\left|\sum_{t=1}^T(\epsilon_t+\delta_t)\right| = |X_T|$, the first term in the cost of the player. In the second term, we note that for each $\alpha\in[0,1]$, the expectation of $\Delta_\alpha = \sum_{t=1}^T (x_t - p_t)\cdot\mathbb{1}[p_t=\alpha]$ is exactly $(1/2-\alpha)$ times the expected number of times α is predicted. If we "assume" that $\mathbb{E}[|\Delta_\alpha|]$ is equal to the absolute value of $\mathbb{E}[\Delta_\alpha]$[7], the second term can be equivalently written as

$$\sum_{\alpha\in[0,1]} |\alpha - 1/2|^2 \cdot \sum_{t=1}^T \mathbb{1}[p_t = \alpha] = \sum_{t=1}^T (p_t - 1/2)^2 = \sum_{t=1}^T \epsilon_t^2.$$

Therefore, an upper bound on the cost in the controlled random walk game gives a uniform upper bound on $\sum_{\alpha\in[0,1]} f(\alpha)\cdot\Delta_\alpha$ over all $f\in\mathcal{F}$ and thus, by definition, upper bounds $\mathsf{smCE}(x,p)$.

A strategy with sub-$\sqrt{T}$ cost. The trivial strategy of playing $\epsilon_t = 0$ at every step gives a cost of $\mathbb{E}[|X_T|] = \mathbb{E}\left[\left|\sum_{t=1}^T \delta_t\right|\right] = \Theta(\sqrt{T})$. Can we do better? Consider the following simple strategy: Fix a parameter $\epsilon\in(0,1/2]$, and play $\epsilon_t = -\epsilon\cdot\operatorname{sgn}(X_{t-1})$ at step t. In other words, we move towards the origin by a distance of ϵ at each step. This strategy clearly gives $\sum_{t=1}^T \epsilon_t^2 \le T\epsilon^2$.

For the $|X_T|$ term, the following heuristic argument suggests $\mathbb{E}[|X_T|] = O(1/\epsilon)$. Assume that the noise δ_t follows the standard Gaussian instead of the uniform distribution on $\{\pm 1/2\}$. Then, the random process $X_t = X_{t-1} - \epsilon\cdot\operatorname{sgn}(X_t) + \delta_t$ is a discretization of the following dynamics:

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = -\nabla U(X(t)) + \mathrm{d}B(t),$$

where the potential is $U(x) = \epsilon|x|$, and $B(t)$ is the standard Brownian motion. As $t\to+\infty$, we expect the distribution of X_t to converge to a distribution with density at x proportional to $e^{-\beta U(x)}$ for some constant β, i.e., the Laplace distribution. This implies $\mathbb{E}[|X_t|] = O(1/\epsilon)$.

If we set $\epsilon = T^{-1/3}$, both terms in the cost would be bounded by $O(T^{1/3})$, which gives a polynomial improvement over the trivial cost of $\sqrt{T}$. Formalizing this heuristic argument gives a forecasting strategy with $\mathbb{E}[\mathsf{smCE}(x,p)] = O(T^{1/3})$ against random bits. Since the number of different predictions (i.e., the number of different values of ϵ_t) is a constant, applying the second part of Theorem 2 shows that $\mathsf{CalDist}(x,p)$ is also at most $O(T^{1/3})$ in expectation.

[7] This is clearly false in general. In the actual proof, we use a concentration argument on Δ_α to show that this approximation is essentially true.

A strategy with polylog(T) **cost.** The improvement from $T^{1/3}$ to polylog(T) is done by varying the parameter ϵ throughout the game. Interestingly, unlike the usual "doubling trick" in online learning, our time horizon is divided into epochs with geometrically *decreasing* lengths.

Suppose that in the first $T/2$ steps, the player simply drifts with the noise. Typically, we expect to have $|X_{T/2}| = O(\sqrt{T})$. For concreteness, assume that $X_{T/2} = \sqrt{T}$. Starting from time $T/2+1$, we play $\epsilon_t = -\alpha/\sqrt{T}$ for some $\alpha = \text{polylog}(T)$. Then, we expect that X_t will drop below 0 before the game ends with high probability. This is because, if the game (hypothetically) runs for $T/2$ more steps, X_T would roughly follow a Gaussian with mean $\sqrt{T} - \frac{T}{2} \cdot \frac{\alpha}{\sqrt{T}} = -\Omega(\alpha\sqrt{T})$ and variance $O(T)$. For sufficiently large α, this will be negative with high probability. Therefore, the player simply waits for X_t to become negative, at which point X_t should be very close to 0. Now, there are at most $T/2$ steps remaining, and we repeat the same strategy for the rest of the game.

This strategy clearly controls X_t such that $|X_T| = O(1)$ with high probability. To upper bound the $\sum_{t=1}^T \epsilon_t^2$ term, note that in the first epoch, we predict a value with absolute value $\alpha/\sqrt{T}$ at most $T/2$ times. This contributes at most $\frac{T}{2} \cdot (\alpha/\sqrt{T})^2 = \text{polylog}(T)$ to the sum. Since we repeat this at most $O(\log T)$ times, we end up with a polylog(T) cost. Finally, since the procedure only involves $O(\log T)$ different values of ϵ_t, applying the second bound in Theorem 2 bounds the calibration distance by polylog(T) as well.

3.4 Calibration Distance Lower Bound

Our lower bound proof is based on the observation that a lower bound for the controlled random walk game also gives a lower bound on the smooth calibration error. Again, we write $p_t = 1/2 - \epsilon_t$ and $x_t = 1/2 + \delta_t$. Then, the location X_T of the player after T steps is exactly given by $X_T = \sum_{t=1}^T (\epsilon_t + \delta_t) = \sum_{t=1}^T (x_t - p_t)$, the difference between the total outcomes and total predictions. Recall that $\mathsf{smCE}(x, p)$ is the supremum of $\sum_{t=1}^T f(p_t) \cdot (x_t - p_t)$ among all 1-Lipschitz functions $f : [0,1] \to [-1,1]$. In particular, by picking f to be the constant function 1 or -1, we obtain $\mathsf{smCE}(x, p) \geq |X_T|$.

To see how the $\sum_{t=1}^T \epsilon_t^2$ term comes into play, consider the function $f(x) = 1/2 - x$, which is in the family $\mathcal{F}$. This gives $\sum_{t=1}^T f(p_t) \cdot (x_t - p_t) = \sum_{t=1}^T \epsilon_t \cdot (\epsilon_t + \delta_t)$. After taking an expectation, the $\epsilon_t \cdot \delta_t$ term vanishes. Therefore, at least in expectation, the smooth calibration error is lower bounded by the $\sum_{t=1}^T \epsilon_t^2$ term as well.

However, as outlined in Section 3.3, the player can achieve a polylog(T) cost in the controlled random walk game. To obtain the $\Omega(T^{1/3})$ lower bound, we make another simple observation: as long as we can lower bound the expectation of $\max_{t \in [T]} |X_t| + \sum_{t=1}^T \epsilon_t^2$ by $\Omega(T^{1/3})$, we can obtain the same lower bound in the prediction setting via an *early stopping* trick, which was used by [QV21] in their lower bound on the ECE.

Indeed, if an algorithm gives $\mathbb{E}\left[\sum_{t=1}^T \epsilon_t^2\right] = \Omega(T^{1/3})$, the connection that we made earlier lower bounds $\mathbb{E}\left[\mathsf{smCE}(x, p)\right]$ by $\Omega(T^{1/3})$ as desired. Otherwise, the value $|X_t|$ must be large at some point t'. Equivalently, the bias $\left|\sum_{t=1}^{t'} (x_t - p_t)\right|$ is large. Then, if the adversary deviates from outputting random bits, and keep outputting the same bit, this large bias will remain in the end.

To lower bound this strengthened cost of $\max_{t \in [T]} |X_t| + \sum_{t=1}^T \epsilon_t^2$, we divide the horizon T into $T^{1/3}$ *epochs* of length $T^{2/3}$ each. In a typical block, the sum of δ_t has an absolute value of $\Omega(\sqrt{T^{2/3}}) = \Omega(T^{1/3})$. Then, if the total control of the player (i.e., the sum of ϵ_t) is much smaller than $T^{1/3}$ in absolute value, we will catch a large $|X_t|$ during this epoch. In order not to be caught,

the player is forced to ensure that the sum of ϵ_t is $\pm\Omega(T^{1/3})$ in a typical epoch. This, in turn, lower bounds the sum of ϵ_t^2 within that epoch by $\Omega(1)$. Summing over the $T^{1/3}$ epochs gives the desired lower bound of $\sum_{t=1}^{T} \epsilon_t^2 = \Omega(T^{1/3})$.

4 Discussion and Open Problems

Is the calibration distance a good metric? In this work, we propose to use the calibration distance as a calibration measure in sequential prediction setups. The definition of the calibration distance is natural, and in the same spirit as the work of [BGHN23] for the offline setup. Compared to the ECE, the calibration distance is better-behaved in being Lipschitz continuous in the predictions. Compared to alternative calibration measures that are continuous (such as weak and smooth calibration), the calibration distance is, from the forecaster's perspective, especially easy and intuitive to certify—to show that the calibration distance is small, the forecaster only needs to output a set of alternative predictions that are calibrated and close to the actual predictions. From this perspective, our proof of Theorem 2 is *algorithmic* in the sense that it implies an efficient algorithm for the forecaster to find such a certificate.

On the other hand, the calibration distance is still far from being the "perfect" calibration measure in the sequential setup. As we highlight in Proposition 1, even on a sequence of random bits, minimizing the calibration distance might incentivize the forecaster to deviate from the "right" predictions. Unfortunately, this is unavoidable to some extent—as discussed in Section 1.2, such incentive-related issues may arise even when only perfect calibration is concerned.

Stronger approximation guarantees. An obvious open problem is to strengthen Theorem 1 and design better approximation algorithms for the calibration distance. A natural avenue is through refining the structural results that relate the calibration distance to the lower calibration distance. More concretely, is the $O(\sqrt{T})$ bound on the gap between $\mathsf{CalDist}(x, p)$ and $\mathsf{LowerCalDist}(x, p)$ tight? Can we avoid the extra $O(1)$ multiplicative factor for the sparse case?

Explicit and efficient algorithms. Our proof of Theorem 3 is based on the minimax theorem and thus non-constructive. Deriving an actual algorithm requires solving a zero-sum game with an action space that is doubly-exponential in T. Is there an explicit and efficient algorithm for matching the $O(\sqrt{T})$ guarantee? A concrete approach is based on the prior work of [KF08, FH18, FH21], which gave *deterministic* forecasters that asymptotically satisfy weak or smooth calibration. Roughly speaking, their forecasting algorithms are deterministic because they are based on fixed point theorems rather than the minimax theorem. While these work focused on asymptotic calibration rather than the exact convergence bounds, it follows easily from [KF08, Lemma 4.3] that their algorithm gives $\mathbb{E}\left[\mathsf{smCE}(x, p)\right] = O(T^{2/3})$ in our notations. Is there a more refined analysis of their approach that gives an $O(\sqrt{T})$ bound? Can we efficiently implement their algorithms, which, as stated, are based on finding fixed points?

Stronger lower bounds. Another obvious open problem is to strengthen the $\Omega(T^{1/3})$ lower bound, which is essentially based a sequence of random bits (each taking value 1 with probability 1/2). A natural attempt is to divide the time horizon into multiple epochs, and use different probabilities for different epochs. The issue is, of course, that the forecaster may predict strategically to decrease the error that it has accumulated in previous epochs. For the ECE, this difficulty was

partially resolved by [QV21] via a "sidestepping" technique, which uses an adaptive, divide-and-conquer strategy for choosing the probabilities for different epochs. Can we use the same approach to bootstrap the $T^{1/3}$ lower bound to a rate closer to $T^{1/2}$?

5 Proof of the Approximation Guarantees

In this section, we prove Theorem 2, which states that the lower calibration distance is a good additive approximation of the calibration distance. Then, Theorem 1 follows easily from the algorithm of [BGHN23] that computes the lower calibration distance up to an ϵ additive error in $\mathrm{poly}(n, 1/\epsilon)$ time.

5.1 Impossibility of Multiplicative Approximation

Before we proceed to the proof, we give a concrete example showing that $\mathsf{LowerCalDist}(x, p)$ does not give a good multiplicative approximation of $\mathsf{CalDist}(x, p)$, even if the multiplicative factor is allowed to depend on T (but not on x or p).

Proposition 4. *There is no function $f : \mathbb{N} \to (0, +\infty)$ such that the following holds for all T, $x \in \{0,1\}^T$ and $p \in [0,1]^T$:*

$$\mathsf{LowerCalDist}(x, p) \geq f(T) \cdot \mathsf{CalDist}(x, p). \tag{2}$$

The proof is based on an example inspired by the proof of [BGHN23, Lemma 4.5].

Proof. Let $T = 4$, $x = (0, 1, 0, 1)$, and $p = (1/2 - \epsilon, 1/2 - \epsilon, 1/2 + \epsilon, 1/2 + \epsilon)$ for some small $\epsilon \in (0, 1/30)$. We will show that $\mathsf{CalDist}(x, p) = 4\epsilon$ while $\mathsf{LowerCalDist}(x, p) = O(\epsilon^2)$. This implies that the ratio

$$\frac{\mathsf{LowerCalDist}(x, p)}{\mathsf{CalDist}(x, p)} \leq \frac{O(\epsilon^2)}{4\epsilon} = O(\epsilon)$$

can be made arbitrarily small by taking $\epsilon \to 0^+$, so Inequality (2) cannot hold for any fixed f.

The calibration distance. We first note that $\mathsf{CalDist}(x, p) \leq 4\epsilon$, as $q = (1/2, 1/2, 1/2, 1/2)$ is perfectly calibrated with respect to x, and $\|p - q\|_1 = 4\epsilon$. Furthermore, for any $q \in \mathcal{C}(x)$, all the entries of q must lie in

$$\{a/b : a \in \{0, 1, \ldots, b\}, b \in [4]\} = \{0, 1/4, 1/3, 1/2, 2/3, 3/4, 1\}.$$

If any entry q_i is different from $1/2$, the difference $|p_i - q_i|$ must be at least $(1/2 - \epsilon) - 1/3 = 1/6 - \epsilon \geq 4\epsilon$ (the last step follows from $\epsilon < 1/30$), which implies $\|p - q\|_1 \geq 4\epsilon$. This shows $\mathsf{CalDist}(x, p) = 4\epsilon$.

The lower calibration distance. Roughly speaking, the $O(\epsilon^2)$ lower calibration distance is achieved by transporting $O(\epsilon)$ units of the "extra ones (resp., zeros)" at $1/2 - \epsilon$ (resp., $1/2 + \epsilon$) to $1/2$. Formally, let $\mathcal{D}_1$ and $\mathcal{D}_4$ be the degenerate distributions supported on $\{1/2 - \epsilon\}$ and $\{1/2 + \epsilon\}$, respectively. Let $\beta = \frac{1/2-\epsilon}{1/2+\epsilon}$, and define $\mathcal{D}_2$ and $\mathcal{D}_3$ as

$$\mathcal{D}_2(1/2 - \epsilon) = \mathcal{D}_3(1/2 + \epsilon) = \beta, \qquad \mathcal{D}_2(1/2) = \mathcal{D}_3(1/2) = 1 - \beta.$$

We can then verify that $(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4) \in \underline{\mathcal{C}}(x)$, because for $\alpha = 1/2$, we have

$$\sum_{t=1}^{T}(x_t - \alpha) \cdot \mathcal{D}_t(\alpha) = (x_2 - 1/2) \cdot (1-\beta) + (x_3 - 1/2) \cdot (1-\beta) = \frac{1-\beta}{2} - \frac{1-\beta}{2} = 0,$$

and for $\alpha = 1/2 - \epsilon$, we have

$$\sum_{t=1}^{T}(x_t - \alpha) \cdot \mathcal{D}_t(\alpha) = (x_1 - 1/2 + \epsilon) \cdot 1 + (x_2 - 1/2 + \epsilon) \cdot \beta = 0.$$

The $\alpha = 1/2 + \epsilon$ case holds by symmetry. By definition of the lower calibration distance,

$$\mathsf{LowerCalDist}(x, p) \le \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \mathcal{D}_t}[|p_t - q_t|] = 0 + (1-\beta) \cdot \epsilon + (1-\beta) \cdot \epsilon + 0 = \frac{4\epsilon^2}{1/2 + \epsilon} = O(\epsilon^2).$$

□

5.2 Rounding of Distributions with a Small Support

We start with the following lemma, which is crucial for proving both bounds in Theorem 2.

Lemma 5. *Suppose that $x \in \{0,1\}^T$, $p \in [0,1]^T$, and $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T) \in \underline{\mathcal{C}}(x)$, where $\mathcal{D}_1, \ldots, \mathcal{D}_T$ are distributions supported over a finite set $S \subset [0,1]$. Then,*

$$\mathsf{CalDist}(x, p) \le \|p - \mathcal{D}\|_1 + 4|S|.$$

The lemma states that if we have distributions $\mathcal{D}_1$ through $\mathcal{D}_T$ supported on a common set of a small size, and they serve as a witness for $\mathsf{LowerCalDist}(x, p)$ being small, we can "round" them to a sequence of (deterministic) predictions and show that $\mathsf{CalDist}(x, p)$ is small as well. The rounding procedure only leads to an additive increase in the distance that is linear in the support size.

Proof. We will first transform $\mathcal{D}_1, \ldots, \mathcal{D}_T$ to another T distributions $(\mathcal{D}'_1, \ldots, \mathcal{D}'_T) \in \underline{\mathcal{C}}(x)$ that satisfy a monotonicity property. Furthermore, the new distributions are still over set S, and the cost $\|p - \mathcal{D}'\|_1$ does not exceed the original cost $\|p - \mathcal{D}\|_1$. With this monotonicity, we apply a simple rounding scheme to produce a good witness $q \in \mathcal{C}(x)$ which shows that $\mathsf{CalDist}(x, p)$ is small.

Enforce monotonicity. For $b \in \{0,1\}$, let $\mathcal{T}^{(b)} \coloneqq \{t \in [T] : x_t = b\}$ denote the set of time steps where the outcome is b. We claim that there exist distributions $\mathcal{D}'_1, \ldots, \mathcal{D}'_T$ over S such that:

- $\mathcal{D}' = (\mathcal{D}'_1, \ldots, \mathcal{D}'_T) \in \underline{\mathcal{C}}(x)$.
- $\|p - \mathcal{D}'\|_1 \le \|p - \mathcal{D}\|_1$.
- There exist total orders on $\mathcal{T}^{(0)}$ and $\mathcal{T}^{(1)}$ (both denoted by "$\prec$") such that: For every $b \in \{0,1\}$ and $t_1, t_2 \in \mathcal{T}^{(b)}$, $t_1 \prec t_2$ implies: (1) $p_{t_1} \le p_{t_2}$; (2) the maximum element in the support of $\mathcal{D}'_{t_1}$ is less than or equal to the minimum element in the support of $\mathcal{D}'_{t_2}$.

In words, the third condition requires that among all the time steps in $\mathcal{T}^{(b)}$, steps with a small value of p_t corresponds to a distribution $\mathcal{D}'_t$ supported over smaller values.

Construction of $\mathcal{D}'$. The existence of $\mathcal{D}'$ should be obvious when viewing the problem as an optimal transport in one dimension: For each $b \in \{0,1\}$, we originally have one unit of mass on p_t for each $t \in \mathcal{T}^{(b)}$, while the goal is to obtain the configuration $\sum_{t\in\mathcal{T}^{(b)}} \mathcal{D}_t$. In order to minimize the total cost, we should match the two measures in increasing order.

Nevertheless, we provide an elementary proof of this claim by explicitly constructing the distributions $\mathcal{D}'_1$ through $\mathcal{D}'_T$. For each $b \in \{0,1\}$, let $m \coloneqq \left|\mathcal{T}^{(b)}\right|$ and $t_1, t_2, \ldots, t_m$ be a permutation of the elements of $\mathcal{T}^{(b)}$ such that $p_{t_1} \le p_{t_2} \le \cdots \le p_{t_m}$. Then, we set $\mathcal{D}'_{t_1}, \mathcal{D}'_{t_2}, \ldots, \mathcal{D}'_{t_m}$ as the unique distributions over set S that satisfy:

- $\sum_{i=1}^m \mathcal{D}'_{t_i} = \sum_{i=1}^m \mathcal{D}_{t_i}$, i.e., for every $\alpha \in S$, $\sum_{i=1}^m \mathcal{D}'_{t_i}(\alpha) = \sum_{i=1}^m \mathcal{D}_{t_i}(\alpha)$.
- For every $i \in [m-1]$, the maximum element in the support of $\mathcal{D}'_{t_i}$ is less than or equal to the minimum element in the support of $\mathcal{D}'_{t_{i+1}}$.

More concretely, these m distributions can be found by starting with the total measure $\sum_{i=1}^m \mathcal{D}_{t_i}$, and then greedily forming a probability measure by taking one unit of mass from the smallest elements in the support of the remaining measure, until m probability measures are formed.

By construction, the distributions $\mathcal{D}'_1, \ldots, \mathcal{D}'_T$ are over set S, and satisfy the monotonicity constraint (with respect to the total order defined as $t_1 \prec t_2 \prec \cdots \prec t_m$). It remains to show that $\mathcal{D}' \in \underline{\mathcal{C}}(x)$ and that the cost of $\mathcal{D}'$ is not higher than that of $\mathcal{D}$.

$\mathcal{D}'$ is perfectly calibrated. We note that for every $\alpha \in [0,1]$,

$$\sum_{t=1}^T (x_t - \alpha)\mathcal{D}'_t(\alpha) = \sum_{b\in\{0,1\}} (b-\alpha) \sum_{t\in\mathcal{T}^{(b)}} \mathcal{D}'_t(\alpha) = \sum_{b\in\{0,1\}} (b-\alpha) \sum_{t\in\mathcal{T}^{(b)}} \mathcal{D}_t(\alpha) = \sum_{t=1}^T (x_t - \alpha)\mathcal{D}_t(\alpha) = 0.$$

The second step follows from our construction of $\mathcal{D}'$, while the last step holds since $\mathcal{D} \in \underline{\mathcal{C}}(x)$. This proves $\mathcal{D}' \in \underline{\mathcal{C}}(x)$.

$\mathcal{D}'$ does not have a higher cost. Fix $b \in \{0,1\}$. Let $a_0 < a_1 < \cdots < a_m$ be the distinct elements in $\{p_1, \ldots, p_T\} \cup S$. For each $i \in [m]$, define

$$F_i \coloneqq \sum_{t\in\mathcal{T}^{(b)}} \mathbb{1}\left[p_t < a_i\right] \quad \text{and} \quad G_i \coloneqq \sum_{t\in\mathcal{T}^{(b)}} \Pr_{q_t\sim\mathcal{D}_t}\left[q_t < a_i\right].$$

We claim that

$$\sum_{t\in\mathcal{T}^{(b)}} \mathop{\mathbb{E}}_{q_t\sim\mathcal{D}'_t}\left[|p_t - q_t|\right] = \sum_{i=1}^m (a_i - a_{i-1}) \cdot |F_i - G_i| \le \sum_{t\in\mathcal{T}^{(b)}} \mathop{\mathbb{E}}_{q_t\sim\mathcal{D}_t}\left[|p_t - q_t|\right]. \tag{3}$$

Summing over $b \in \{0,1\}$ proves $\|p - \mathcal{D}'\|_1 \le \|p - \mathcal{D}\|_1$.

We first prove the first step in Equation (3). For each $i \in [m]$, let δ_i denote the amount of mass transported across the interval $[a_{i-1}, a_i]$ according to $\mathcal{D}'$. Formally, we define

$$\delta_i \coloneqq \sum_{t\in\mathcal{T}^{(b)}} \Pr_{q_t\sim\mathcal{D}'_t}\left[[a_{i-1}, a_i] \subseteq [\min\{p_t, q_t\}, \max\{p_t, q_t\}]\right].$$

For any $p, q \in \{a_0, a_1, \ldots, a_m\}$, we have the identity

$$|p - q| = \sum_{i=1}^{m} (a_i - a_{i-1}) \cdot \mathbb{1}\left[[a_{i-1}, a_i] \subseteq [\min\{p, q\}, \max\{p, q\}]\right].$$

Thus, we can re-write the cost $\sum_{t \in \mathcal{T}^{(b)}} \mathbb{E}_{q_t \sim \mathcal{D}'_t}\left[|p_t - q_t|\right]$ as $\sum_{i=1}^{m} (a_i - a_{i-1}) \cdot \delta_i$, and it remains to prove that $\delta_i = |F_i - G_i|$ for every $i \in [m]$.

Fix $i \in [m]$. Let $t_1 \prec t_2 \prec \cdots$ be the elements of $\mathcal{T}^{(b)}$ sorted according to total order $\prec$. Recall that F_i is the number of values among $\{p_t : t \in T^{(b)}\}$ that are strictly smaller than a_i, so we have $p_{t_j} \le a_{i-1}$ for every $j \le F_i$ and $p_{t_j} \ge a_i$ for every $j > F_i$.

Suppose that $F_i \ge G_i$. By construction of $\mathcal{D}'_t$, for every $j \le \lfloor G_i \rfloor$, the support of $\mathcal{D}'_{t_j}$ is contained in $[0, a_{i-1}]$ while $p_{t_j} \in [0, a_{i-1}]$, so they do not contribute to δ_i. For $j \in \{\lceil G_i \rceil + 1, \ldots, F_i\}$, the support of $\mathcal{D}'_{t_j}$ is completely contained in $[a_i, 1]$ while $p_{t_j} \in [0, a_{i-1}]$, so they contribute $F_i - \lceil G_i \rceil$ to δ_i. Finally, when G_i is not integral, for $j = \lceil G_i \rceil$, we have $p_{t_j} \in [0, a_{i-1}]$ and $\mathcal{D}'_{t_j}$ assigns a probability mass of $\lceil G_i \rceil - G_i$ to $[a_i, 1]$. This contributes $\lceil G_i \rceil - G_i$ to δ_i. Therefore, we conclude that, in this case,

$$\delta_i = (F_i - \lceil G_i \rceil) + (\lceil G_i \rceil - G_i) = |F_i - G_i|.$$

The $F_i < G_i$ case is similar. For $j \le F_i$, we have $p_{t_j} \in [0, a_{i-1}]$, while the support of $\mathcal{D}'_{t_j}$ is also contained in $[0, a_{i-1}]$, so these values of j do not contribute to δ_i. When $F_i + 1 \le j \le \lfloor G_i \rfloor$, the support of $\mathcal{D}'_{t_j}$ is still contained in $[0, a_{i-1}]$, whereas $p_{t_j} \ge a_i$. This contributes $\lfloor G_i \rfloor - F_i$ to δ_i. Finally, when G_i is not integral, for $j = \lceil G_i \rceil$, $\mathcal{D}'_{t_j}$ assigns a probability mass of $G_i - \lfloor G_i \rfloor$ to $[0, a_{i-1}]$, and this contributes $G_i - \lfloor G_i \rfloor$ to δ_i. Again, we have $\delta_i = G_i - F_i = |F_i - G_i|$.

Next, we prove the second step in Equation (3), i.e., $\sum_{t \in \mathcal{T}^{(b)}} \mathbb{E}_{q_t \sim \mathcal{D}_t}\left[|p_t - q_t|\right]$ is lower bounded by $\sum_{i=1}^{m} (a_i - a_{i-1}) \cdot |F_i - G_i|$. Similarly, we define

$$\delta_i \coloneqq \sum_{t \in \mathcal{T}^{(b)}} \Pr_{q_t \sim \mathcal{D}_t}\left[[a_{i-1}, a_i] \subseteq [\min\{p_t, q_t\}, \max\{p_t, q_t\}]\right]$$

as the total mass transported across $[a_{i-1}, a_i]$ according to $\mathcal{D}$, and it suffices to show that $\delta_i \ge |F_i - G_i|$ for every $i \in [m]$.

Fix $i \in [m]$ and suppose that $F_i \ge G_i$. Note that for any $p, q \in \{a_0, a_1, \ldots, a_m\}$, we have the inequality

$$\mathbb{1}\left[p \le a_{i-1} \wedge q \ge a_i\right] \ge \mathbb{1}\left[p < a_i\right] - \mathbb{1}\left[q < a_i\right].$$

Then, by definition of δ_i, we have

$$\delta_i \ge \sum_{t \in \mathcal{T}^{(b)}} \Pr_{q_t \sim \mathcal{D}_t}\left[p_t \le a_{i-1} \wedge q_t \ge a_i\right] \ge \sum_{t \in \mathcal{T}^{(b)}} \mathbb{1}\left[p_t < a_i\right] - \sum_{t \in \mathcal{T}^{(b)}} \Pr_{q_t \sim \mathcal{D}_t}\left[q_t < a_i\right] = |F_i - G_i|.$$

Similarly, when $F_i < G_i$, using the inequality

$$\mathbb{1}\left[p \ge a_i \wedge q \le a_{i-1}\right] \ge \mathbb{1}\left[p \ge a_i\right] - \mathbb{1}\left[q \ge a_i\right],$$

we get

$$\begin{aligned}\delta_i \ge \sum_{t \in \mathcal{T}^{(b)}} \Pr_{q_t \sim \mathcal{D}_t}\left[p_t \ge a_i \wedge q_t \le a_{i-1}\right] &\ge \sum_{t \in \mathcal{T}^{(b)}} \mathbb{1}\left[p_t \ge a_i\right] - \sum_{t \in \mathcal{T}^{(b)}} \Pr_{q_t \sim \mathcal{D}_t}\left[q_t \ge a_i\right] \\ &= \left(\left|\mathcal{T}^{(b)}\right| - F_i\right) - \left(\left|\mathcal{T}^{(b)}\right| - G_i\right) = |F_i - G_i|.\end{aligned}$$

This concludes the proof of the inequality $\|p - \mathcal{D}'\|_1 \le \|p - \mathcal{D}\|_1$, and shows that $\mathcal{D}'$ indeed has all the desired properties.

The rounding scheme. Now that the distributions $\mathcal{D}'_1, \ldots, \mathcal{D}'_T$ have all the nice properties, it remains to find $q \in \mathcal{C}(x)$ such that $\|p - q\|_1 \le \|p - \mathcal{D}'\|_1 + 4|S|$, since the lemma would then follow from

$$\mathsf{CalDist}(x, p) \le \|p - q\|_1 \le \|p - \mathcal{D}'\|_1 + 4|S| \le \|p - \mathcal{D}\|_1 + 4|S|.$$

Let $\mathcal{T}^{\mathsf{pure}}$ be the set of indices $t \in [T]$ such that $\mathcal{D}'_t$ is degenerate (i.e., with a size-1 support). We call each $t \in \mathcal{T}^{\mathsf{pure}}$ a "pure step", and each $t \in \mathcal{T}^{\mathsf{mixed}} \coloneqq [T] \setminus \mathcal{T}^{\mathsf{pure}}$ a "mixed step". For each pure step $t \in \mathcal{T}^{\mathsf{pure}}$, let $\beta_t \in [0, 1]$ be the (only) element in the support of $\mathcal{D}'_t$.

We choose $q \in [0, 1]^T$ as follows:

- For each pure step $t \in \mathcal{T}^{\mathsf{pure}}$, set q_t to
$$g(\beta_t) \coloneqq \frac{\sum_{t' \in \mathcal{T}^{\mathsf{pure}}} x_{t'} \cdot \mathbb{1}\left[\beta_{t'} = \beta_t\right]}{\sum_{t' \in \mathcal{T}^{\mathsf{pure}}} \mathbb{1}\left[\beta_{t'} = \beta_t\right]}.$$
- For each mixed step $t \in \mathcal{T}^{\mathsf{mixed}}$, set $q_t = x_t$.

In the remainder of the proof, we verify that $q \in \mathcal{C}(x)$ and then upper bound $\|p - q\|_1$.

q is perfectly calibrated. For every $\alpha \in [0, 1]$, we can write

$$\sum_{t=1}^{T} (x_t - q_t) \cdot \mathbb{1}\left[q_t = \alpha\right] = \sum_{t \in \mathcal{T}^{\mathsf{pure}}} (x_t - q_t) \cdot \mathbb{1}\left[q_t = \alpha\right] + \sum_{t \in \mathcal{T}^{\mathsf{mixed}}} (x_t - q_t) \cdot \mathbb{1}\left[q_t = \alpha\right].$$

The second summation is 0, since $q_t = x_t$ holds for every $t \in \mathcal{T}^{\mathsf{mixed}}$. By our choice of q, the first summation can be written as

$$\begin{aligned}
&\sum_{t \in \mathcal{T}^{\mathsf{pure}}} (x_t - q_t) \cdot \mathbb{1}\left[g(\beta_t) = \alpha\right] \\
&= \sum_{\alpha' \in S} \mathbb{1}\left[g(\alpha') = \alpha\right] \sum_{t \in \mathcal{T}^{\mathsf{pure}}} (x_t - \alpha) \cdot \mathbb{1}\left[\beta_t = \alpha'\right] \\
&= \sum_{\alpha' \in S} \mathbb{1}\left[g(\alpha') = \alpha\right] \cdot \left[\sum_{t \in \mathcal{T}^{\mathsf{pure}}} x_t \cdot \mathbb{1}\left[\beta_t = \alpha'\right] - g(\alpha') \cdot \sum_{t \in \mathcal{T}^{\mathsf{pure}}} \mathbb{1}\left[\beta_t = \alpha'\right]\right] \\
&= 0,
\end{aligned}$$

where the last step follows from the definition of $g(\cdot)$. This verifies $q \in \mathcal{C}(x)$.

Upper bound $\|p - q\|_1$. Note that

$$\begin{aligned}
\|p - q\|_1 &= \sum_{t \in \mathcal{T}^{\mathsf{pure}}} |p_t - q_t| + \sum_{t \in \mathcal{T}^{\mathsf{mixed}}} |p_t - q_t| \\
&\le \sum_{t \in \mathcal{T}^{\mathsf{pure}}} |p_t - \beta_t| + \sum_{t \in \mathcal{T}^{\mathsf{pure}}} |q_t - \beta_t| + \left|\mathcal{T}^{\mathsf{mixed}}\right| \\
&\le \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \mathcal{D}'_t} \left[|p_t - q_t|\right] + \sum_{t \in \mathcal{T}^{\mathsf{pure}}} |q_t - \beta_t| + \left|\mathcal{T}^{\mathsf{mixed}}\right|.
\end{aligned}$$

In the following, we will show that both $\sum_{t \in \mathcal{T}^{\mathsf{pure}}} |q_t - \beta_t|$ and $\left|\mathcal{T}^{\mathsf{mixed}}\right|$ are upper bounded by $2|S|$, which would conclude the proof.

Bound the number of mixed steps. We start by showing $|\mathcal{T}^{\mathsf{mixed}}| \le 2|S|$. Fix $t \in \mathcal{T}^{\mathsf{mixed}}$. Let $b = x_t$ and $s \in S$ be the largest element in the support of $\mathcal{D}'_t$. Since $\mathcal{D}'_t$ is not degenerate, the support of $\mathcal{D}'_t$ contains another element $s' < s$. Recall that $\mathcal{T}^{(b)}$ is associated with a total order $\prec$ that is consistent with both p_t's and the supports of $\mathcal{D}'_t$. Then, with respect to order $\prec$, t must be the smallest index in $\mathcal{T}^{(b)}$ such that $\mathcal{D}'_t(s) \neq 0$. Indeed, if there exists $t' \prec t$ such that $\mathcal{D}'_{t'}(s) > 0$, the fact that the support of $\mathcal{D}'_t$ contains a smaller element $s' < s$ contradicts the monotonicity. Therefore, we showed that every $t \in \mathcal{T}^{\mathsf{mixed}}$ corresponds to a unique pair $(b, s) \in \{0, 1\} \times S$. This implies $|\mathcal{T}^{\mathsf{mixed}}| \le 2|S|$.

Bound the additional cost on the pure steps. Fix $s \in S$. For each $b \in \{0, 1\}$, let

$$n_b \coloneqq \sum_{t \in \mathcal{T}^{\mathsf{pure}}} \mathbb{1}\left[\beta_t = s \wedge x_t = b\right] \quad \text{and} \quad \epsilon_b \coloneqq \sum_{t \in \mathcal{T}^{\mathsf{mixed}}} \mathcal{D}'_t(s) \cdot \mathbb{1}\left[x_t = b\right].$$

We claim that

$$s = \frac{n_1 + \epsilon_1}{n_0 + n_1 + \epsilon_0 + \epsilon_1}$$

and

$$g(s) = \frac{n_1}{n_0 + n_1}.$$

The latter follows immediately from the definition of $g(\cdot)$, n_0 and n_1. The former holds since $\mathcal{D}' \in \underline{\mathcal{C}}(x)$ implies

$$\begin{aligned} 0 &= \sum_{t=1}^{T} (x_t - s) \cdot \mathcal{D}'_t(s) \\ &= \sum_{t \in \mathcal{T}^{\mathsf{pure}}} (x_t - s) \cdot \mathcal{D}'_t(s) + \sum_{t \in \mathcal{T}^{\mathsf{mixed}}} (x_t - s) \cdot \mathcal{D}'_t(s) \\ &= [n_1 - s \cdot (n_0 + n_1)] + [\epsilon_1 - s \cdot (\epsilon_0 + \epsilon_1)], \end{aligned}$$

which, after rearrangement, gives the expression of s.

We also argue that $\epsilon_0, \epsilon_1 \in [0, 2]$. Fix $b \in \{0, 1\}$, and let $t_1, t_2 \in \mathcal{T}^{(b)}$ be the smallest and the largest index $t \in \mathcal{T}^{(b)}$ (with respect to total order $\prec$) such that $\mathcal{D}'_t(s) \neq 0$. By monotonicity, for any $t \in \mathcal{T}^{(b)}$ such that $t_1 \prec t \prec t_2$, the support of $\mathcal{D}'_t$ can only contain s, which impies $t \in \mathcal{T}^{\mathsf{pure}}$. Therefore, ϵ_b is exactly given by $\sum_{t \in \{t_1, t_2\}} \mathcal{D}'_t(s)$, which clearly lies in $[0, 2]$.

Then, we have

$$(n_0 + n_1) \cdot |s - g(s)| = (n_0 + n_1) \cdot \left| \frac{n_1 + \epsilon_1}{n_0 + n_1 + \epsilon_0 + \epsilon_1} - \frac{n_1}{n_0 + n_1} \right|.$$

For fixed n_0 and n_1, the last expression is maximized when either $(\epsilon_0, \epsilon_1) = (2, 0)$ or $(\epsilon_0, \epsilon_1) = (0, 2)$. A simple calculation shows that the expression is upper bounded by 2 in both cases.

Finally, we note that

$$\sum_{t \in \mathcal{T}^{\mathsf{pure}}} |q_t - \beta_t| = \sum_{s \in S} \sum_{t \in \mathcal{T}^{\mathsf{pure}}} |q_t - \beta_t| \cdot \mathbb{1}\left[\beta_t = s\right] = \sum_{s \in S} |g(s) - s| \cdot \sum_{t \in \mathcal{T}^{\mathsf{pure}}} \mathbb{1}\left[\beta_t = s\right].$$

Therefore, the contribution of each $s \in S$ to $\sum_{t \in \mathcal{T}^{\mathsf{pure}}} |q_t - \beta_t|$ is exactly $(n_0 + n_1) \cdot |s - g(s)| \le 2$. Thus, we have $\sum_{t \in \mathcal{T}^{\mathsf{pure}}} |q_t - \beta_t| \le 2|S|$. □

5.3 Proof of the Additive Gap

With Lemma 5 in hand, we prove the first part of Theorem 2, which upper bounds the gap between $\mathsf{CalDist}(x,p)$ and $\mathsf{LowerCalDist}(x,p)$ by $O(\sqrt{T})$.

The proof starts by finding T distributions $\hat{\mathcal{D}}_1,\dots,\hat{\mathcal{D}}_T$ that (approximately) achieve the lower calibration distance $\mathsf{LowerCalDist}(x,p)$. We refine these distributions to $\mathcal{D}_1,\dots,\mathcal{D}_T$, so that: (1) the support of every $\mathcal{D}_t$ is contained in the same set of size $O(\sqrt{T})$; (2) $\mathcal{D}$ still approximately achieves $\mathsf{LowerCalDist}(x,p)$ up to an $O(\sqrt{T})$ slack. This allows us to invoke our rounding lemma (Lemma 5) to show $\mathsf{CalDist}(x,p) \le \mathsf{LowerCalDist}(x,p) + O(\sqrt{T})$.

Before proceeding with the proof below, it would be helpful to recall the connection between the lower calibration distance and optimal transport (Remark 2). In particular, during the proof we will sometimes refer to the distributions $(\mathcal{D}_1,\dots,\mathcal{D}_T) \in \underline{\mathcal{C}}(x)$ and the corresponding transportation of the bits interchangeably.

Lemma 6. *For any $x \in \{0,1\}^T$, $p \in [0,1]^T$, there exists a set $S \subseteq [0,1]$ of size at most $O(\sqrt{T})$ along with distributions $\mathcal{D}_1,\dots,\mathcal{D}_T$ over S, such that $(\mathcal{D}_1,\mathcal{D}_2,\dots,\mathcal{D}_T) \in \underline{\mathcal{C}}(x)$ and*

$$\|p - \mathcal{D}\|_1 \le \mathsf{LowerCalDist}(x,p) + O(\sqrt{T}).$$

Proof. By definition of the lower calibration distance, there exists $\hat{\mathcal{D}} = (\hat{\mathcal{D}}_1,\hat{\mathcal{D}}_2,\dots,\hat{\mathcal{D}}_T) \in \underline{\mathcal{C}}(x)$ such that $\|p - \hat{\mathcal{D}}\|_1 \le \mathsf{LowerCalDist}(x,p) + 1$.[8]

Pick an integer $m = \Theta(\sqrt{T})$ and define intervals $\mathcal{I}_i := [(i-1)/m, i/m)$ for $i \in [m-1]$ and $\mathcal{I}_m := [(m-1)/m, 1]$. Roughly speaking, for each $i \in [m]$, we will examine the bits that are transported into the interval $\mathcal{I}_i$ according to $\hat{\mathcal{D}}$. We then consolidate these bits into a single destination. The perfect calibration of $\hat{\mathcal{D}}$ implies that the new destination still falls into $\mathcal{I}_i$, so our change increases the cost by at most $T/m = O(\sqrt{T})$. Furthermore, the new transportation only involves at most m different destinations (one for each interval $\mathcal{I}_i$).

The new destinations. For each $i \in [m]$, we define

$$\mu_i := \frac{\sum_{t=1}^T x_t \cdot \hat{\mathcal{D}}_t(\mathcal{I}_i)}{\sum_{t=1}^T \hat{\mathcal{D}}_t(\mathcal{I}_i)} = \frac{\sum_{\alpha\in\mathcal{I}_i}\sum_{t=1}^T x_t \cdot \hat{\mathcal{D}}_t(\alpha)}{\sum_{\alpha\in\mathcal{I}_i}\sum_{t=1}^T \hat{\mathcal{D}}_t(\alpha)}.$$

Note that by definition of $\underline{\mathcal{C}}(x)$, each $\hat{\mathcal{D}}_t$ has a finite support, so both $\sum_{t=1}^T x_t \cdot \hat{\mathcal{D}}_t(\alpha)$ and $\sum_{t=1}^T \hat{\mathcal{D}}_t(\alpha)$ take non-zero values only on finitely many choices of α. Therefore, the summations over α in the last expression above are actually finite and thus well-defined.

We argue that for every $i \in [m]$, μ_i falls into $[(i-1)/m, i/m]$. Indeed, we can re-write μ_i as

$$\mu_i = \sum_{\alpha\in\mathcal{I}_i} \frac{\sum_{t=1}^T x_t \cdot \hat{\mathcal{D}}_t(\alpha)}{\sum_{\beta\in\mathcal{I}_i}\sum_{t=1}^T \hat{\mathcal{D}}_t(\beta)} = \sum_{\alpha\in\mathcal{I}_i} \frac{\sum_{t=1}^T \hat{\mathcal{D}}_t(\alpha)}{\sum_{\beta\in\mathcal{I}_i}\sum_{t=1}^T \hat{\mathcal{D}}_t(\beta)} \cdot \frac{\sum_{t=1}^T x_t \cdot \hat{\mathcal{D}}_t(\alpha)}{\sum_{t=1}^T \hat{\mathcal{D}}_t(\alpha)}. \tag{4}$$

For each $\alpha \in \mathcal{I}_i$, it follows from $\hat{\mathcal{D}} \in \underline{\mathcal{C}}(x)$ that $\sum_{t=1}^T (x_t - \alpha)\cdot \hat{\mathcal{D}}_t(\alpha) = 0$ and, equivalently, $\frac{\sum_{t=1}^T x_t \cdot \hat{\mathcal{D}}_t(\alpha)}{\sum_{t=1}^T \hat{\mathcal{D}}_t(\alpha)} = \alpha \in \mathcal{I}_i$. Therefore, Equation (4) states that μ_i is a convex combination of values that lie in interval $\mathcal{I}_i$, which ensures that $(i-1)/m \le \mu_i \le i/m$.

[8] We need the "+1" term in case that the infimum in the definition of $\mathsf{LowerCalDist}(x,p)$ cannot be achieved.

The updated transportation. We define another T distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T$ as follows. Let $\phi : [0,1] \to [0,1]$ be the function that maps every value in each $\mathcal{I}_i$ to μ_i. Then, each $\mathcal{D}_t$ is defined as the distribution of $\phi(q_t)$ when $q_t \sim \hat{\mathcal{D}}_t$. We will argue that $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_T)$ is in $\underline{\mathcal{C}}(x)$ and that $\|p - \mathcal{D}\|_1$ is comparable to the cost $\|p - \hat{\mathcal{D}}\|_1$.

To show that $\mathcal{D} \in \underline{\mathcal{C}}(x)$, we note that for any $\alpha \in [0,1]$, we have

$$\begin{aligned}
\sum_{t=1}^{T}(x_t - \alpha) \cdot \mathcal{D}_t(\alpha) &= \sum_{t=1}^{T}(x_t - \alpha) \cdot \Pr_{q_t \sim \hat{\mathcal{D}}_t}[\phi(q_t) = \alpha] && \text{(definition of } \mathcal{D}_t\text{)} \\
&= \sum_{t=1}^{T}(x_t - \alpha) \cdot \sum_{i=1}^{m} \mathbb{1}[\mu_i = \alpha] \cdot \hat{\mathcal{D}}_t(\mathcal{I}_i) && \text{(definition of } \phi\text{)} \\
&= \sum_{i=1}^{m} \mathbb{1}[\mu_i = \alpha] \cdot \left[\sum_{t=1}^{T} x_t \cdot \hat{\mathcal{D}}_t(\mathcal{I}_i) - \mu_i \cdot \sum_{t=1}^{T} \hat{\mathcal{D}}_t(\mathcal{I}_i)\right] = 0. && \text{(definition of } \mu_i\text{)}
\end{aligned}$$

This proves $\mathcal{D} \in \underline{\mathcal{C}}(x)$.

To show the latter property, we note that since $\phi(x)$ and x always fall into the same interval $\mathcal{I}_i$, we have $|\phi(x) - x| \le 1/m$ for every $x \in [0,1]$. Then,

$$\begin{aligned}
\sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \mathcal{D}_t}[|p_t - q_t|] &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t}[|p_t - \phi(q_t)|] && \text{(definition of } \mathcal{D}_t\text{)} \\
&\le \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t}[|p_t - q_t|] + \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t}[|q_t - \phi(q_t)|] && \text{(triangle inequality)} \\
&\le [\mathsf{LowerCalDist}(x,p) + 1] + T \cdot \frac{1}{m} && \text{(choice of } \hat{\mathcal{D}} \text{ and } |x - \phi(x)| \le 1/m\text{)} \\
&\le \mathsf{LowerCalDist}(x,p) + O(\sqrt{T}). && (m = \Theta(\sqrt{T}))
\end{aligned}$$

Finally, note that each $\mathcal{D}_t$ is over the same set of size $\le m = O(\sqrt{T})$, namely, $\{\mu_1, \mu_2, \ldots, \mu_m\}$. This concludes the proof. □

The first bound in Theorem 2 then follows easily.

Proof of the first part of Theorem 2. By Lemma 6, there exists $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T) \in \underline{\mathcal{C}}(x)$ such that $\|p - \mathcal{D}\|_1 \le \mathsf{LowerCalDist}(x,p) + O(\sqrt{T})$, and each $\mathcal{D}_t$ is over the same set of size $O(\sqrt{T})$. By Lemma 5, we have

$$\mathsf{CalDist}(x,p) \le \|p - \mathcal{D}\|_1 + O(\sqrt{T}) \le \mathsf{LowerCalDist}(x,p) + O(\sqrt{T}).$$

□

5.4 Approximation Guarantee in the Sparse Case

Now we deal with the second part of Theorem 2, where we have a prediction sequence $p \in [0,1]^T$ with only m different entries. In order to invoke our Lemma 5, however, we need to show that the lower calibration distance $\mathsf{LowerCalDist}(x,p)$ can be approximately achieved by distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ over a small set $S \subset [0,1]$ (more concretely, of size $O(m)$). This step, stated as the lemma below, turns out to be much more complicated.

Lemma 7. *For any $x \in \{0,1\}^T$, $p \in [0,1]^T$, $m = |\{p_1, p_2, \ldots, p_T\}|$ and $\epsilon > 0$, there exists a set $S \subset [0,1]$ of size at most $2m+3$ along with distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ over S, such that*

$$(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T) \in \underline{\mathcal{C}}(x)$$

and

$$\|p - \mathcal{D}\|_1 \leq 20 \cdot \mathsf{LowerCalDist}(x,p) + \epsilon.$$

We first show how Lemmas 5 and 7 together imply the second part of Theorem 2.

Proof of the second part of Theorem 2. Applying Lemma 7 with $\epsilon = 1$ gives a set $S \subset [0,1]$ of size $\leq 2m+3$, together with distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ over S such that $\mathcal{D} \in \underline{\mathcal{C}}(x)$ and

$$\|p - \mathcal{D}\|_1 \leq 20 \cdot \mathsf{LowerCalDist}(x,p) + 1.$$

Then, by Lemma 5,

$$\mathsf{CalDist}(x,p) \leq \|p - \mathcal{D}\|_1 + 4 \cdot (2m+3) \leq 20 \cdot \mathsf{LowerCalDist}(x,p) + (8m+13).$$

□

Before we prove Lemma 7, again, we recommend the reader to review Remark 2. In the following proof, we will frequently mention the transportation of the bits in lieu of the explicit expressions for the probability distributions.

As in the proof of the first part of Theorem 2, we start by picking $\hat{\mathcal{D}} \in \underline{\mathcal{C}}(x)$ that approximately achieves $\mathsf{LowerCalDist}(x,p)$, and then consolidate the transportation specified by $\hat{\mathcal{D}}$ such that there will be at most $O(m)$ destinations.

Let $s_1 < s_2 < \cdots < s_m$ be the m values that appear in the entries of p. A natural first attempt would be to examine the bits that are transported into each interval $[s_i, s_{i+1}]$ and merge them to a single destination. Unfortunately, as we show in Appendix B, this naïve consolidation could blow up the cost. Instead, our proof of Lemma 7 involves a much more complicated case analysis based on the amounts of zeros and ones being transported into $[s_i, s_{i+1}]$ on both directions.

Proof of Lemma 7. By definition of $\mathsf{LowerCalDist}(x,p)$, there exist T distributions $\hat{\mathcal{D}}_1, \hat{\mathcal{D}}_2, \ldots, \hat{\mathcal{D}}_T$ such that:

- Each $\hat{\mathcal{D}}_t$ is supported over a finite subset of $[0,1]$.
- $\hat{\mathcal{D}}_1, \ldots, \hat{\mathcal{D}}_T$ are perfectly calibrated, i.e., $\sum_{t=1}^{T} (x_t - \alpha) \cdot \hat{\mathcal{D}}_t(\alpha) = 0$ holds for every $\alpha \in [0,1]$.
- The cost is close to $\mathsf{LowerCalDist}(x,p)$: $\|p - \hat{\mathcal{D}}\|_1 \leq \mathsf{LowerCalDist}(x,p) + \epsilon/20$.[9]

[9] Again, we need the "$\epsilon/20$" term because the infimum in the definition of $\mathsf{LowerCalDist}(x,p)$ might not be achieved by any $\mathcal{D} \in \underline{\mathcal{C}}(x)$.

Proof overview. Let $0 = s_1 < s_2 < \cdots < s_{m'} = 1$ be the distinct values among $\{p_1, p_2, \ldots, p_T\} \cup \{0, 1\}$. Note that $m' \le |\{p_1, p_2, \ldots, p_T\}| + 2 = m + 2$. We will transform $\hat{\mathcal{D}}_1, \ldots, \hat{\mathcal{D}}_T$ into another T distributions, denoted by $\mathcal{D}_1, \ldots, \mathcal{D}_T$, over a set $S \subset [0, 1]$, such that:

- $|S| \le 2m' - 1 \le 2m + 3$;
- $(\mathcal{D}_1, \ldots, \mathcal{D}_T) \in \underline{\mathcal{C}}(x)$;
- $\|p - \mathcal{D}\|_1 \le 20 \cdot \|p - \hat{\mathcal{D}}\|_1$.

Note that doing so would prove the lemma, since the last condition implies that

$$\|p - \mathcal{D}\|_1 \le 20\|p - \hat{\mathcal{D}}\|_1 \le 20 \cdot (\mathsf{LowerCalDist}(x, p) + \epsilon/20) = 20 \cdot \mathsf{LowerCalDist}(x, p) + \epsilon.$$

To ensure the first property, we examine the probability masses that $\hat{\mathcal{D}}_1$ through $\hat{\mathcal{D}}_T$ assign to the interval (s_i, s_{i+1}) for each $i \in [m' - 1]$. These can be interpreted as a way of transporting certain fractions of the bits $x_1, \ldots, x_T$ to the interval, so that the resulting configuration is calibrated. A priori, the bits might be transported to many different destinations within the interval (s_i, s_{i+1}). We will reroute the transportation, so that the bits will only arrive at five different destinations: 0, 1, s_i, s_{i+1}, and another unique value assigned for this interval. In the end, the supports of $\mathcal{D}_1$ through $\mathcal{D}_T$ will be among $s_1, s_2, \ldots, s_{m'}$ along with $m' - 1$ other values. Therefore, the corresponding set S will have size at most $2m' - 1$.

Decomposition of costs. Our first step is to decompose the cost $\|p - \hat{\mathcal{D}}\|_1$ into a few parts. Let $\mathcal{I}_i \coloneqq [s_i, s_{i+1})$ for every $i \in [m' - 2]$ and $\mathcal{I}_{m'-1} \coloneqq [s_{m'-1}, s_{m'}] = [s_{m'-1}, 1]$. The total cost associated with interval $\mathcal{I}_i$ is defined as

$$\mathsf{Cost}_i \coloneqq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[|p_t - q_t| \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \right] \right].$$

Furthermore, for each interval $\mathcal{I}_i$, we decompose the cost according to whether the transportation is from the left or from the right (we view the interval $[0, 1]$ as a line segment in which the small values lie on the left):

$$\mathsf{Cost}_i^{\mathsf{L}} \coloneqq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[|p_t - q_t| \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \le s_i \right] \right],$$
$$\mathsf{Cost}_i^{\mathsf{R}} \coloneqq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[|p_t - q_t| \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \ge s_{i+1} \right] \right].$$

Finally, note that whenever the condition $q_t \in \mathcal{I}_i \wedge p_t \le s_i$ holds in the definition of $\mathsf{Cost}_i^{\mathsf{L}}$, we have $p_t \le s_i \le q_t$, which gives $|p_t - q_t| = |p_t - s_i| + |q_t - s_i|$. Therefore, we can write $\mathsf{Cost}_i^{\mathsf{L}} = \mathsf{Cost}_i^{\mathsf{L,out}} + \mathsf{Cost}_i^{\mathsf{L,in}}$, where

$$\mathsf{Cost}_i^{\mathsf{L,out}} \coloneqq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[|p_t - s_i| \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \le s_i \right] \right],$$
$$\mathsf{Cost}_i^{\mathsf{L,in}} \coloneqq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[|q_t - s_i| \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \le s_i \right] \right].$$

Here the superscripts "out" and "in" specify whether the cost is for the transportation outside or inside the interval $\mathcal{I}_i$. Similarly, we decompose $\mathsf{Cost}_i^{\mathsf{R}}$ into the following two terms:

$$\mathsf{Cost}_i^{\mathsf{R},\mathsf{out}} \coloneqq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[|p_t - s_{i+1}| \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \geq s_{i+1} \right] \right],$$
$$\mathsf{Cost}_i^{\mathsf{R},\mathsf{in}} \coloneqq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[|q_t - s_{i+1}| \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \geq s_{i+1} \right] \right].$$

Our use of the word "decompose" can be justified by the following identity:

$$\begin{aligned}
\sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[|p_t - q_t| \right] &= \sum_{i=1}^{m'-1} \mathsf{Cost}_i \\
&= \sum_{i=1}^{m'-1} (\mathsf{Cost}_i^{\mathsf{L}} + \mathsf{Cost}_i^{\mathsf{R}}) \\
&= \sum_{i=1}^{m'-1} (\mathsf{Cost}_i^{\mathsf{L},\mathsf{out}} + \mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{out}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{in}}).
\end{aligned}$$

The first step holds since $\mathcal{I}_1, \ldots, \mathcal{I}_{m'-1}$ form a partition of $[0,1]$, which implies $1 = \sum_{i=1}^{m'-1} \mathbb{1}\left[x \in \mathcal{I}_i \right]$ for any $x \in [0,1]$. The second step follows from the observation that p_t never falls into (s_i, s_{i+1}), so we have $\mathbb{1}\left[q_t \in \mathcal{I}_i \right] = \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \leq s_i \right] + \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \geq s_{i+1} \right]$.

Lower bound the cost of the second phase. For $b \in \{0,1\}$, let $\mathsf{unit}_{i,b}^{\mathsf{L}}$ denote the amount of bit b that is transported into interval $\mathcal{I}_i$ from $[0, s_i]$. Formally,

$$\mathsf{unit}_{i,b}^{\mathsf{L}} \coloneqq \sum_{t=1}^{T} \hat{\mathcal{D}}_t(\mathcal{I}_i) \cdot \mathbb{1}\left[x_t = b \wedge p_t \leq s_i \right].$$

Similarly, $\mathsf{unit}_{i,b}^{\mathsf{R}}$ is defined as the amount of bit b moved from $[s_{i+1}, 1]$ to $\mathcal{I}_i$:

$$\mathsf{unit}_{i,b}^{\mathsf{R}} \coloneqq \sum_{t=1}^{T} \hat{\mathcal{D}}_t(\mathcal{I}_i) \cdot \mathbb{1}\left[x_t = b \wedge p_t \geq s_{i+1} \right].$$

Intuitively, $\hat{\mathcal{D}}_1$ through $\hat{\mathcal{D}}_T$ specify the following transportation of bits:

- For each $i \in [m'-1]$, we spend a total cost of $\mathsf{Cost}_i^{\mathsf{L},\mathsf{out}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{out}}$ to transport zeros and ones from $[0, s_i] \cup [s_{i+1}, 1]$ to either s_i and s_{i+1} ("the first phase").
- At this point, there are $\mathsf{unit}_{i,0}^{\mathsf{L}}$ (resp., $\mathsf{unit}_{i,1}^{\mathsf{L}}$) units of zeros (resp., ones) at s_i, and $\mathsf{unit}_{i,b}^{\mathsf{R}}$ units of bit b at s_{i+1}.
- Then, we further distribute these bits to values within $\mathcal{I}_i$ so that the outcomes are calibrated ("the second phase"), at a total cost of $\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{in}}$.

The distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ that we will define is based on a new transportation that keeps the total cost of the first phase (outside of $\mathcal{I}_i$). We will change the second phase, so that there will be at most one destination outside $\{s_1, s_2, \ldots, s_{m'}\}$. Furthermore, we make sure that this change only blows up the cost of the second phase by a constant factor.

For this purpose, we start by lower bounding $\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{in}}$. Let $\Delta_i^{\mathsf{L}} \coloneqq \mathsf{unit}_{i,1}^{\mathsf{L}} - (\mathsf{unit}_{i,0}^{\mathsf{L}} + \mathsf{unit}_{i,1}^{\mathsf{L}}) \cdot s_i$ and $\Delta_i^{\mathsf{R}} \coloneqq \mathsf{unit}_{i,1}^{\mathsf{R}} - (\mathsf{unit}_{i,0}^{\mathsf{R}} + \mathsf{unit}_{i,1}^{\mathsf{R}}) \cdot s_{i+1}$ denote the biases incurred at point s_i and s_{i+1} between the first and the second phases. We will prove the following inequality: For any 1-Lipschitz function $f : [0,1] \to [-1,1]$,

$$2\left(\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{in}}\right) \geq f(s_i) \cdot \Delta_i^{\mathsf{L}} + f(s_{i+1}) \cdot \Delta_i^{\mathsf{R}}. \tag{5}$$

The following proof is the same as the one in [BGHN23] for lower bounding the lower distance from calibration by the smooth calibration error. We include the proof for completeness.

Fix a 1-Lipschitz function $f : [0,1] \to [-1,1]$. Consider the function $g_b(v) \coloneqq f(v) \cdot (b - v)$ defined over $[0,1]$ for $b \in \{0,1\}$. Since $|g_b'(v)| = |f'(v)(b-v) - f(v)| \leq 2$ for any $v \in [0,1]$, g_b is 2-Lipschitz. Then, we have

$$\begin{aligned}
2\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[2|q_t - s_i| \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \leq s_i\right]\right] \\
&\geq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[\left[f(s_i) \cdot (x_t - s_i) - f(q_t) \cdot (x_t - q_t)\right] \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \leq s_i\right]\right] \\
&\qquad\qquad (v \mapsto f(v) \cdot (x_t - v) \text{ is 2-Lipschitz}) \\
&= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[f(s_i) \cdot (x_t - s_i) \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \leq s_i\right]\right] \\
&\quad - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[f(q_t) \cdot (x_t - q_t) \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \leq s_i\right]\right].
\end{aligned}$$

The first summation in the last expression above can be further simplified into:

$$\begin{aligned}
&f(s_i) \cdot \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[x_t \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \leq s_i\right]\right] - f(s_i) \cdot s_i \cdot \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[\mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \leq s_i\right]\right] \\
= f(s_i) &\cdot \sum_{t=1}^{T} \hat{\mathcal{D}}_t(\mathcal{I}_i) \cdot \mathbb{1}\left[x_t = 1 \wedge p_t \leq s_i\right] - f(s_i) \cdot s_i \cdot \sum_{t=1}^{T} \hat{\mathcal{D}}_t(\mathcal{I}_i) \cdot \mathbb{1}\left[p_t \leq s_i\right] \\
= f(s_i) &\cdot \left[\mathsf{unit}_{i,1}^{\mathsf{L}} - s_i \cdot (\mathsf{unit}_{i,0}^{\mathsf{L}} + \mathsf{unit}_{i,1}^{\mathsf{L}})\right] = f(s_i) \cdot \Delta_i^{\mathsf{L}},
\end{aligned}$$

and thus,

$$2\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} \geq f(s_i) \cdot \Delta_i^{\mathsf{L}} - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[f(q_t) \cdot (x_t - q_t) \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \leq s_i\right]\right]. \tag{6}$$

An analogous argument gives

$$2\mathsf{Cost}_i^{\mathsf{R},\mathsf{in}} \geq f(s_{i+1}) \cdot \Delta_i^{\mathsf{R}} - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[f(q_t) \cdot (x_t - q_t) \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i \wedge p_t \geq s_{i+1}\right]\right]. \tag{7}$$

Finally, Inequality (5) follows from Inequalities (6) and (7), together with the observation that the two summations on the right-hand sides of (6) and (7) sum up to

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[f(q_t) \cdot (x_t - q_t) \cdot \mathbb{1}\left[q_t \in \mathcal{I}_i\right] \right] = \sum_{\alpha \in \mathcal{I}_i} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \hat{\mathcal{D}}_t} \left[f(q_t) \cdot (x_t - q_t) \cdot \mathbb{1}\left[q_t = \alpha\right] \right]$$
$$= \sum_{\alpha \in \mathcal{I}_i} f(\alpha) \cdot \sum_{t=1}^{T} (x_t - \alpha) \cdot \hat{\mathcal{D}}_t(\alpha) = 0,$$

where the last step follows from $\hat{D} \in \underline{\mathcal{C}}(x)$.

Given Inequality (5), we apply Lemma 20 from Appendix B to lower bound $\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{in}}$ by a closed-form expression of Δ_i^{L}, Δ_i^{R}, s_i, and s_{i+1}.

$$2\left(\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{in}}\right) \geq \begin{cases} |\Delta_i^{\mathsf{L}}| + |\Delta_i^{\mathsf{R}}|, & \Delta_i^{\mathsf{L}} \cdot \Delta_i^{\mathsf{R}} \geq 0, \\ |\Delta_i^{\mathsf{L}} + \Delta_i^{\mathsf{R}}| + (s_{i+1} - s_i) \cdot \min\{|\Delta_i^{\mathsf{L}}|, |\Delta_i^{\mathsf{R}}|\}, & \Delta_i^{\mathsf{L}} \cdot \Delta_i^{\mathsf{R}} < 0. \end{cases} \tag{8}$$

Handle the same-sign situation. It remains to change the second phase of the transportation inside interval $\mathcal{I}_i$, so that there will be at most one destination (in addition to $s_1, s_2, \ldots, s_{m'}$), while the cost is bounded by $20 \cdot (\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{in}})$.

We start by noting that we may assume $\min\left\{\mathsf{unit}_{i,0}^{\mathsf{L}}, \mathsf{unit}_{i,1}^{\mathsf{L}}\right\} = \min\left\{\mathsf{unit}_{i,0}^{\mathsf{R}}, \mathsf{unit}_{i,1}^{\mathsf{R}}\right\} = 0$ without loss of generality. This is because, for example, when both $\mathsf{unit}_{i,0}^{\mathsf{L}}$ and $\mathsf{unit}_{i,1}^{\mathsf{L}}$ are positive, we may take $\mu \coloneqq \min\{\mathsf{unit}_{i,0}^{\mathsf{L}}/(1-s_i), \mathsf{unit}_{i,1}^{\mathsf{L}}/s_i\}$, and let $\mu \cdot s_i$ units of ones and $\mu \cdot (1-s_i)$ units of zeros be "settled" at s_i. After this, either $\mathsf{unit}_{i,0}^{\mathsf{L}}$ or $\mathsf{unit}_{i,1}^{\mathsf{L}}$ becomes zero, and the quantity Δ_i^{L} is unchanged. The same argument applies to $\mathsf{unit}_{i,0}^{\mathsf{R}}$ and $\mathsf{unit}_{i,1}^{\mathsf{R}}$ as well.

We first deal with the case that $\Delta_i^{\mathsf{L}}, \Delta_i^{\mathsf{R}} \geq 0$. In this case, we have $\mathsf{unit}_{i,0}^{\mathsf{L}} = \mathsf{unit}_{i,0}^{\mathsf{R}} = 0$, i.e., there are no extra zeros at either s_i or s_{i+1}, though there might be extra ones. We will transport these ones to 1, at a cost of

$$\mathsf{unit}_{i,1}^{\mathsf{L}} \cdot (1 - s_i) + \mathsf{unit}_{i,1}^{\mathsf{R}} \cdot (1 - s_{i+1}) = \Delta_i^{\mathsf{L}} + \Delta_i^{\mathsf{R}} = |\Delta_i^{\mathsf{L}}| + |\Delta_i^{\mathsf{R}}| \leq 2\left(\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{in}}\right).$$

The last step above follows from Equation (8). Similarly, if $\Delta_i^{\mathsf{L}}, \Delta_i^{\mathsf{R}} \leq 0$, we have $\mathsf{unit}_{i,1}^{\mathsf{L}} = \mathsf{unit}_{i,1}^{\mathsf{R}} = 0$. We will transport all the extra zeros to 0, and the total cost will be

$$\mathsf{unit}_{i,0}^{\mathsf{L}} \cdot s_i + \mathsf{unit}_{i,0}^{\mathsf{R}} \cdot s_{i+1} = -\Delta_i^{\mathsf{L}} - \Delta_i^{\mathsf{R}} = |\Delta_i^{\mathsf{L}}| + |\Delta_i^{\mathsf{R}}| \leq 2\left(\mathsf{Cost}_i^{\mathsf{L},\mathsf{in}} + \mathsf{Cost}_i^{\mathsf{R},\mathsf{in}}\right).$$

In both cases, we settle all the bits that were originally associated with interval $\mathcal{I}_i$ at a total cost of at most $2\mathsf{Cost}_i$. Furthermore, all the destinations lie in the set $\{0, s_i, s_{i+1}, 1\}$.

Handling opposite signs, the first part. The case that $\Delta_i^{\mathsf{L}} \cdot \Delta_i^{\mathsf{R}} < 0$ is more involved. We first deal with the case that $\Delta_i^{\mathsf{L}} > 0$ and $\Delta_i^{\mathsf{R}} < 0$. Recall that we assumed $\min\{\mathsf{unit}_{i,0}^{\mathsf{L}}, \mathsf{unit}_{i,1}^{\mathsf{L}}\} = \min\{\mathsf{unit}_{i,0}^{\mathsf{R}}, \mathsf{unit}_{i,1}^{\mathsf{R}}\} = 0$ without loss of generality. This means that $\mathsf{unit}_{i,1}^{\mathsf{L}}, \mathsf{unit}_{i,0}^{\mathsf{R}} > 0$, while $\mathsf{unit}_{i,0}^{\mathsf{L}} = \mathsf{unit}_{i,1}^{\mathsf{R}} = 0$.

We shorthand $x \coloneqq \mathsf{unit}_{i,1}^{\mathsf{L}}$ and $y \coloneqq \mathsf{unit}_{i,0}^{\mathsf{R}}$. Our strategy is to move all the bits—the x units of ones at s_i and the y units of zeros at s_{i+1}—to value $p \coloneqq \frac{x}{x+y}$. The total cost would be

$$x \cdot |p - s_i| + y \cdot |p - s_{i+1}| = (x + y) \cdot \left[p \cdot |p - s_i| + (1 - p) \cdot |p - s_{i+1}|\right].$$

Note that $\Delta_i^{\mathsf{L}} = x \cdot (1 - s_i)$ and $\Delta_i^{\mathsf{R}} = -y \cdot s_{i+1}$. The right-hand side of Inequality (8) can then be re-written as

$$\begin{aligned}&|x \cdot (1 - s_i) - y \cdot s_{i+1}| + (s_{i+1} - s_i) \cdot \min\{x \cdot (1 - s_i), y \cdot s_{i+1}\} \\ &= (x + y) \cdot [|p \cdot (1 - s_i) - (1 - p) \cdot s_{i+1}| + (s_{i+1} - s_i) \cdot \min\{p \cdot (1 - s_i), (1 - p) \cdot s_{i+1}\}] .\end{aligned}$$

Then, applying Lemma 21 from Appendix B with $\alpha = s_i$ and $\beta = s_{i+1}$ shows that the cost of the new transportation is at most

$$2 \cdot \left[|\Delta_i^{\mathsf{L}} + \Delta_i^{\mathsf{R}}| + (s_{i+1} - s_i) \cdot \min\{|\Delta_i^{\mathsf{L}}|, |\Delta_i^{\mathsf{R}}|\}\right] \le 4 \cdot \left(\mathsf{Cost}_i^{\mathsf{L,in}} + \mathsf{Cost}_i^{\mathsf{R,in}}\right).$$

Handling opposite signs, the second part. It remains to handle the case that $\Delta_i^{\mathsf{L}} < 0$ and $\Delta_i^{\mathsf{R}} > 0$. In this case, we have $\mathsf{unit}_{i,0}^{\mathsf{L}}, \mathsf{unit}_{i,1}^{\mathsf{R}} > 0$, while $\mathsf{unit}_{i,1}^{\mathsf{L}} = \mathsf{unit}_{i,0}^{\mathsf{R}} = 0$.

Again, shorthand $x \coloneqq \mathsf{unit}_{i,0}^{\mathsf{L}}$ and $y \coloneqq \mathsf{unit}_{i,1}^{\mathsf{R}}$. The key difference is that we will consider the following two strategies, and use the one with a lower cost:

- **Strategy 1:** Again, move all the bits—the x units of zeros at s_i and the y units of ones at s_{i+1}—to value $p \coloneqq \frac{y}{x+y}$. The total cost would be

$$x \cdot \left|\frac{y}{x+y} - s_i\right| + y \cdot \left|\frac{y}{x+y} - s_{i+1}\right| = (x + y) \cdot [(1 - p) \cdot |p - s_i| + p \cdot |p - s_{i+1}|] .$$

- **Strategy 2:** Move all the zeros at s_i to 0, and all the ones at s_{i+1} to 1. The total cost is

$$x \cdot s_i + y \cdot (1 - s_{i+1}) = (x + y) \cdot [(1 - p) \cdot s_i + p \cdot (1 - s_{i+1})] .$$

In this case, $\Delta_i^{\mathsf{L}} = -x \cdot s_i$, $\Delta_i^{\mathsf{R}} = y \cdot (1 - s_{i+1})$, and the right-hand side of Inequality (8) is given by

$$\begin{aligned}&|x \cdot s_i - y \cdot (1 - s_{i+1})| + (s_{i+1} - s_i) \cdot \min\{x \cdot s_i, y \cdot (1 - s_{i+1})\} \\ &= (x + y) \cdot [|(1 - p) \cdot s_i - p \cdot (1 - s_{i+1})| + (s_{i+1} - s_i) \cdot \min\{(1 - p) \cdot s_i, p \cdot (1 - s_{i+1})\}] .\end{aligned}$$

We apply Lemma 22 from Appendix B with $\alpha = s_i$ and $\beta = s_{i+1}$ to show that the cost of the new transportation is at most

$$10 \cdot \left[|\Delta_i^{\mathsf{L}} + \Delta_i^{\mathsf{R}}| + (s_{i+1} - s_i) \cdot \min\{|\Delta_i^{\mathsf{L}}|, |\Delta_i^{\mathsf{R}}|\}\right] \le 20 \cdot \left(\mathsf{Cost}_i^{\mathsf{L,in}} + \mathsf{Cost}_i^{\mathsf{R,in}}\right).$$

□

6 Proof of the Upper Bound

We prove Theorem 3 in this section. We first note that it is sufficient to give an algorithm that achieves an $O(\sqrt{T})$ smooth calibration error, since this would imply the desired upper bound as follows:

$$\begin{aligned}\mathbb{E}\left[\mathsf{CalDist}(x, p)\right] &\le \mathbb{E}\left[\mathsf{LowerCalDist}(x, p)\right] + O(\sqrt{T}) && \text{(Theorem 2)} \\ &\le 2\,\mathbb{E}\left[\mathsf{smCE}(x, p)\right] + O(\sqrt{T}) && \text{(Lemma 3)} \\ &\le O(\sqrt{T}).\end{aligned}$$

Our approach is based on a minimax argument similar to that of Hart [Har22] for upper bounding the ECE in sequential calibration. Suppose we already know the adversary's strategy, which might be adaptive and randomized. At each step t, based on the previous outcomes $x_1, \ldots, x_{t-1}$ and predictions $p_1, \ldots, p_{t-1}$, we can calculate the conditional probability of $x_t = 1$. The natural strategy is then to predict this value exactly. Then, we may the view the sequences $x \in \{0,1\}^T$ and $p \in [0,1]^T$ as generated as below:

- At each step t, p_t is adversarially chosen based on $x_{1:(t-1)}$ and $p_{1:(t-1)}$.
- Then, we draw $x_t \sim \mathsf{Bernoulli}(p_t)$.

Recall that the smooth calibration error $\mathsf{smCE}(x, p)$ is defined as

$$\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} f(p_t) \cdot (x_t - p_t),$$

where $\mathcal{F}$ is the family of 1-Lipschitz functions from $[0,1]$ to $[-1,1]$. For each *fixed* $f \in \mathcal{F}$, the random process $(X_0, X_1, \ldots, X_T)$ defined as $X_t \coloneqq \sum_{t'=1}^{t} f(p_{t'}) \cdot (x_{t'} - p_{t'})$ is a martingale with bounded differences, so $X_T = \sum_{t=1}^{T} f(p_t) \cdot (x_t - p_t)$ is bounded by $O(\sqrt{T})$ with high probability. The difficulty, however, is to show that the same upper bound holds even if we take a supremum over all functions $f \in \mathcal{F}$.

6.1 An Online Learning Setting

Our proof is based on viewing the discussion above as an instance of online learning. In particular, we follow a formulation in [RST15a].

An "adversary" and a "player" play a game with T steps. At each step $t \in [T]$, the following happen in sequential order:

- The adversary picks an "instance" $p_t^* \in [0,1]$.
- The player, knowing p_t^*, commits to a distribution $\mathcal{D}_t$ over $[-1,1]$, from which the predicted label $\hat{y}_t$ will be drawn.
- The adversary, knowing $\mathcal{D}_t$, generates the true label $y_t \in [-1,1]$.
- The player draws prediction $\hat{y}_t \sim \mathcal{D}_t$ and incurs a loss of $\ell(\hat{y}_t, y_t)$.

The player's objective is to minimize the *cumulative regret*, defined as the difference between the player's total loss and the total loss incurred by the best hypothesis in hindsight:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t)\right] - \mathbb{E}\left[\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(p_t^*), y_t)\right].$$

The adversary aims to maximize this regret. This setup exactly matches the learning setting defined in [RST15a, Equation (10)].

6.2 Regret Bound and Sequential Rademacher Complexity

The work of [RST15a] gives an upper bound on the optimal regret in the above online learning setting in terms of the *sequential Rademacher complexity* of the function class $\mathcal{F}$.

Definition 2 (Sequential Rademacher Complexity)**.** *The sequential Rademacher complexity of a family $\mathcal{F}$ of functions over $[0,1]$ is*

$$\mathcal{SR}_T(\mathcal{F}) \coloneqq \sup_{z_1,\ldots,z_T} \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^T} \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \sigma_t f(z_t(\sigma_1,\sigma_2,\ldots,\sigma_{t-1})) \right],$$

where the outer supremum is taken over all $(z_1,\ldots,z_T)$ such that each $z_t : \{\pm1\}^{t-1} \to [0,1]$.

Theorem 5 (Theorem 8 of [RST15a])**.** *Suppose that for any $y \in [-1,1]$, the loss function $\ell(\cdot, y)$ is convex and L-Lipschitz. Then, the optimal regret is at most $2L \cdot \mathcal{SR}_T(\mathcal{F})$.*

Finally, we will use the following result that upper bounds $\mathcal{SR}_T(\mathcal{F})$ in terms of the covering numbers of $\mathcal{F}$.

Theorem 6 (Theorem 4 of [RST15b])**.** *Let $\mathcal{F}$ be a family of functions over $[0,1]$. With respect to $z = (z_1, z_2, \ldots, z_T)$ where $z_t : \{\pm1\}^{t-1} \to [0,1]$, a family $\mathcal{F}'$ is a δ-cover of $\mathcal{F}$ if, for any $\sigma \in \{\pm1\}^T$ and $f \in \mathcal{F}$, there exists $f' \in \mathcal{F}'$ such that*

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T} (f(z_t(\sigma_{1:(t-1)})) - f'(z_t(\sigma_{1:(t-1)})))^2} \le \delta.$$

Let $\mathcal{N}(\delta, \mathcal{F}, z)$ denote the size of the smallest δ-cover of $\mathcal{F}$ with respect to z. Then,

$$\mathcal{SR}_T(\mathcal{F}) \le \sup_{z_1,z_2,\ldots,z_T} \inf_{\alpha\in[0,1]} \left\{ 4\alpha T + 12\sqrt{T} \cdot \int_{\alpha}^{1} \sqrt{\log \mathcal{N}(\delta, \mathcal{F}, z)}\, \mathrm{d}\delta \right\}.$$

We apply Theorem 6 to upper bound the sequential Rademacher complexity of the class of Lipschitz functions.

Lemma 8. *Let $\mathcal{F}$ be the family of 1-Lipschitz functions from $[0,1]$ to $[-1,1]$. Then,*

$$\mathcal{SR}_T(\mathcal{F}) = O(\sqrt{T}).$$

The lemma is proved by a standard construction of covers for the class of Lipschitz functions.

Proof. By Theorem 6, $\mathcal{SR}_T(\mathcal{F})$ is upper bounded by

$$\sup_{z_1,z_2,\ldots,z_T} \inf_{\alpha\in[0,1]} \left\{ 4\alpha T + 12\sqrt{T} \cdot \int_{\alpha}^{1} \sqrt{\log \mathcal{N}(\delta, \mathcal{F}, z)}\, \mathrm{d}\delta \right\}. \tag{9}$$

We fix $z = (z_1, \ldots, z_T)$ and $\delta \in (0,1]$, and give an upper bound on $\mathcal{N}(\delta, \mathcal{F}, z)$. Let $k = \lceil 2/\delta \rceil$. For a function $f \in \mathcal{F}$, we construct another function $\hat{f}$ that takes value $\frac{\lfloor f(i/k)\cdot k \rfloor}{k}$ at i/k for each $i \in \{0,1,\ldots,k\}$. On each interval $[(i-1)/k, i/k]$, $\hat{f}$ is the linear interpolation between $f((i-1)/k)$ and $f(i/k)$.

We first note that $\hat{f}$ is 1-Lipschitz. For each $i \in [k]$, since f is 1-Lipschitz, we have

$$\left| f\left(\frac{i-1}{k}\right) \cdot k - f\left(\frac{i}{k}\right) \cdot k \right| = k \cdot \left| f\left(\frac{i-1}{k}\right) - f\left(\frac{i}{k}\right) \right| \le k \cdot \frac{1}{k} = 1.$$

It follows that $\lfloor f((i-1)/k) \cdot k \rfloor$ and $\lfloor f(i/k) \cdot k \rfloor$ differ by at most 1, which implies $|\hat{f}((i-1)/k) - \hat{f}(i/k)| \le 1/k$. Thus, the linear interpolation on the interval $[(i-1)/k, i/k]$ has a slope between ± 1. This shows that $\hat{f}$ is 1-Lipschitz.

Then, we argue that $\hat{f}$ is point-wise close to f. For each $i \in \{0, 1, \ldots, k\}$, we have

$$\left| \hat{f}(i/k) - f(i/k) \right| \le \frac{1}{k} \le \frac{\delta}{2}.$$

For general $x \in [0, 1]$, there exists $i \in \{0, 1, \ldots, k\}$ such that $|x - i/k| \le 1/(2k)$. It follows that

$$\begin{aligned} |\hat{f}(x) - f(x)| &\le |\hat{f}(x) - \hat{f}(i/k)| + |\hat{f}(i/k) - f(i/k)| + |f(i/k) - f(x)| \\ &\le |x - i/k| + \frac{\delta}{2} + |x - i/k| && (f \text{ and } \hat{f} \text{ are 1-Lipschitz}) \\ &\le \frac{1}{2k} + \frac{\delta}{2} + \frac{1}{2k} \le \delta. \end{aligned}$$

In particular, regardless of the value of $\sigma \in \{\pm 1\}^T$, we have

$$\sqrt{\frac{1}{T} \sum_{t=1}^{T} (f(z_t(\sigma_{1:(t-1)})) - \hat{f}(z_t(\sigma_{1:(t-1)})))^2} \le \sqrt{\frac{1}{T} \cdot T\delta^2} = \delta.$$

Then, we show that the resulting function $\hat{f}$ falls into a small set. $\hat{f}$ is uniquely determined by its value on $\{0, 1/k, 2/k, \ldots, 1\}$. $\hat{f}(0)$ is one of the $2k+1$ multiples of $1/k$ in $[-1, 1]$. For each $i \in [k]$, $\hat{f}(i/k) - \hat{f}((i-1)/k)$ falls into $\{-1/k, 0, 1/k\}$. Therefore, $\hat{f}$ falls into a set of size at most

$$(2k+1) \cdot 3^k \le 3^{2k} \le 3^{6/\delta}.$$

This gives $\mathcal{N}(\delta, \mathcal{F}, p) \le 3^{6/\delta}$.

Finally, picking $\alpha = 0$ in the expression in (9) shows that

$$\mathcal{SR}(\mathcal{F}) \le 12\sqrt{T} \int_0^1 \sqrt{\frac{6 \log 3}{\delta}} \, \mathrm{d}\delta = O(\sqrt{T}).$$

□

6.3 Proof of Theorem 3

Now we proceed to the proof. A technical issue with the discussion earlier in this section is that, to apply the minimax theorem, the action spaces of the two players need to be finite. This is not true since the forecaster is allowed to make arbitrary predictions between 0 and 1. To deal with this issue, we will force the forecaster to restrict its predictions to a $1/T$-net of $[0, 1]$. Since the smooth calibration error $\mathsf{smCE}(x, p)$ is continuous in p, this rounding does not blow up the error by much.

Proof of Theorem 3. We will show that, even if the forecaster is only allowed to predict the values in $P \coloneqq \{0, 1/T, 2/T, \dots, 1\}$, it is still possible to achieve an $O(\sqrt{T})$ smooth calibration error. By Theorem 2 and Lemma 3, this would give the desired $O(\sqrt{T})$ bound on the calibration distance.

After restricting the space of predictions, a *deterministic* strategy of the adversary (resp., the forecaster) is simply a function from $\bigcup_{t=1}^{T}(\{0,1\} \times P)^{t-1}$ to $\{0,1\}$ (resp., to P). Both sets are finite (albeit of a doubly exponential size). In general, both players may play a mixture of deterministic strategies. By von Neumann's minimax theorem, it suffices to prove that for every fixed (possibly mixed) strategy of the adversary, the forecaster can achieve an $O(\sqrt{T})$ smooth calibration error.

The forecaster's algorithm. Now, we describe one such strategy for the forecaster:

- At each step $t \in [T]$, based on $x_{1:(t-1)}$ and $p_{1:(t-1)}$, compute the conditional probability for the adversary to play $x_t = 1$. Let p_t^* denote this value.
- Predict $p_t \coloneqq \frac{\lfloor T \cdot p_t^* \rfloor}{T}$, which is p_t^* rounded down to the nearest value in P.

We note that it is, in turn, sufficient to upper bound the expected value of $\mathsf{smCE}(p^*, x)$. This is because, by Lemma 19 from Appendix A,

$$\mathbb{E}\left[\mathsf{smCE}(x, p)\right] \le \mathbb{E}\left[\mathsf{smCE}(x, p^*)\right] + 2\,\mathbb{E}\left[\|p - p^*\|_1\right],$$

whereas by our choice of p, $\|p - p^*\|_1$ is always at most $\frac{1}{T} \cdot T = 1$.

After fixing the forecaster's strategy, the "game" between the adversary and the forecaster can be equivalently described as the following procedure. At the beginning, a function g from $\bigcup_{t=1}^{T}(\{0,1\}^{t-1} \times [0,1]^{t-1})$ to $[0,1]$ is adversarially chosen. Then, for $t = 1, 2, \dots, T$:

- Pick $p_t^* = g\left(x_{1:(t-1)}, p^*_{1:(t-1)}\right)$.
- Draw x_t from $\mathsf{Bernoulli}(p_t^*)$.

Note that the first step above is equivalent to the original game, since the predictions $p_{1:(t-1)}$ are determined by $p^*_{1:(t-1)}$.

Reduction to online learning. Now, we further rephrase the procedure described above as an online learning setup in Section 6.1. At each step $t \in [T]$, the adversary picks the "instance" p_t^* as $g\left(x_{1:(t-1)}, p^*_{1:(t-1)}\right)$. The player then commits to an arbitrary distribution $\mathcal{D}_t$ over $[-1, 1]$. (We will show later that the choice of $\mathcal{D}_t$ is inconsequential.) The adversary picks the true label y_t by drawing $x_t \sim \mathsf{Bernoulli}(p_t^*)$ and setting $y_t = x_t - p_t^*$. Finally, the player draws $\hat{y}_t \sim \mathcal{D}_t$ and incurs a loss of $\ell(\hat{y}_t, y_t) \coloneqq \hat{y}_t \cdot y_t$.

The regret in the above setup can be simplified into

$$\mathbb{E}\left[\sum_{t=1}^{T} \hat{y}_t \cdot (x_t - p_t^*)\right] - \mathbb{E}\left[\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f(p_t^*) \cdot (x_t - p_t^*)\right]. \tag{10}$$

By the definition of the learning procedure and the choice of x_t, we have

$$\mathbb{E}\left[\hat{y}_t \cdot (x_t - p_t^*)\right] = \mathop{\mathbb{E}}_{p_t^*, \hat{y}_t}\left[\hat{y}_t \cdot \left(\mathbb{E}\left[x_t | p_t^*, \hat{y}_t\right] - p_t^*\right)\right] = \mathop{\mathbb{E}}_{p_t^*, \hat{y}_t}\left[\hat{y}_t \cdot (p_t^* - p_t^*)\right] = 0$$

for every $t \in [T]$, so the first term in (10) is always 0 regardless of how the player picks $\mathcal{D}_t$ (which determines $\hat{y}_t$). Note that $f \in \mathcal{F}$ if and only if $-f \in \mathcal{F}$, so the expression in (10) is equal to

$$-\mathbb{E}\left[\inf_{f\in\mathcal{F}} \sum_{t=1}^{T} f(p_t^*)\cdot(x_t - p_t^*)\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} f(p_t^*)\cdot(x_t - p_t^*)\right] = \mathbb{E}\left[\mathsf{smCE}(x, p^*)\right].$$

So far, we proved that in our online learning setting, when the adversary plays a specific strategy (namely, pick the instance p_t^* as $g(x_{1:(t-1)}, p_{1:(t-1)})$ and the true label y_t as $x_t - p_t^*$, where $x_t \sim \mathsf{Bernoulli}(p^*)$), the regret of any player is given by $\mathbb{E}\left[\mathsf{smCE}(x, p^*)\right]$. Therefore, $\mathbb{E}\left[\mathsf{smCE}(x, p^*)\right]$ is upper bounded by the optimal regret for this setup. Note that for any $y_t \in [-1, 1]$, the loss function $\ell(\cdot, y_t)$ is convex and 1-Lipschitz in the first parameter. By Theorem 5 and Lemma 8, the optimal regret is at most $2\mathcal{SR}(\mathcal{F}) = O(\sqrt{T})$. This concludes the proof. □

7 Improved Forecasters for Random Bits

In this section, we prove Proposition 1, which gives a polylog(T) calibration distance when the adversary plays T independent random bits. We will first present a simple forecasting algorithm with an $O(T^{1/3})$ calibration distance in expectation. Then, we use the same idea to further improve the calibration distance to $O(\log^{3/2} T)$.

7.1 A Sub-Square-Root Upper Bound for Random Bits

Algorithm 1 gives a forecasting strategy that achieves an $O(T^{1/3})$ smooth calibration error on a sequence of T random bits.

Algorithm 1: Fixed-Bias Forecaster for Random Bits

Input: Time horizon T. Parameter $\epsilon > 0$. Online access to $x_1, x_2, \ldots, x_T$.
Output: Predictions $p_1, p_2, \ldots, p_T$.

$S_0 \leftarrow 0$;
for $t \in [T]$ **do**
  $p_t \leftarrow 1/2 + \epsilon \cdot \mathrm{sgn}(S_{t-1})$;
  Predict p_t;
  Observe x_t;
  $S_t \leftarrow S_{t-1} + (x_t - p_t)$;
end

The algorithm keeps track of $S_t = \sum_{t'=1}^{t}(x_{t'} - p_{t'})$, the difference between the total outcomes and the total predictions in the first t steps. If $S_{t-1} > 0$ (resp., $S_{t-1} < 0$), the forecaster predicts a value slightly higher (resp., lower) than $1/2$, in the hope that S_t will get closer to 0.

Now we analyze the deviation from calibration incurred by the above algorithm. We will start by upper bounding the smooth calibration error of Algorithm 1, and then invoke Theorem 2 and Lemma 3 to get an upper bound on the calibration distance.

We start with the following simple upper bound on the smooth calibration error. Recall the definition of smCE and Δ_α from Section 2.

Lemma 9.

$$\mathsf{smCE}(x,p) \le \left|\sum_{\alpha\in[0,1]} \Delta_\alpha\right| + \sum_{\alpha\in[0,1]} |\alpha - 1/2| \cdot |\Delta_\alpha|.$$

Proof. By definition, we have

$$\begin{aligned}
\mathsf{smCE}(x,p) &= \sup_{f\in\mathcal{F}} \sum_{\alpha\in[0,1]} f(\alpha)\cdot\Delta_\alpha \\
&= \sup_{f\in\mathcal{F}} \left[f(1/2)\cdot \sum_{\alpha\in[0,1]} \Delta_\alpha + \sum_{\alpha\in[0,1]} (f(\alpha)-f(1/2))\cdot\Delta_\alpha \right] \\
&\le \sup_{f\in\mathcal{F}} \left[f(1/2)\cdot \sum_{\alpha\in[0,1]} \Delta_\alpha \right] + \sum_{\alpha\in[0,1]} \sup_{f\in\mathcal{F}} \left[(f(\alpha)-f(1/2))\cdot\Delta_\alpha \right] \\
&\le \left|\sum_{\alpha\in[0,1]} \Delta_\alpha\right| + \sum_{\alpha\in[0,1]} |\alpha-1/2|\cdot|\Delta_\alpha|.
\end{aligned}$$

The last step holds since the functions in $\mathcal{F}$ are both bounded and 1-Lipschitz. □

Note that the $\sum_{\alpha\in[0,1]}\Delta_\alpha$ term in Lemma 9 is exactly $S_T = \sum_{t=1}^{T}(x_t - p_t)$ in Algorithm 1. The following lemma gives a bound on the stochastic process $(S_0, S_1, \ldots, S_T)$.

Lemma 10. *For $\epsilon \in (0, \frac{1}{2}]$, consider the stochastic process $(X_0, X_1, X_2, \ldots)$ defined as follows:*

- $X_0 = 0$.
- $x_1, x_2, x_3, \ldots$ *are independent samples from* $\mathsf{Bernoulli}(1/2)$.
- *For $t \ge 1$, $X_t = X_{t-1} + x_t - \left(\frac{1}{2} + \epsilon\cdot\mathrm{sgn}(X_{t-1})\right)$.*

Then, for any $t \ge 0$, $\Delta \ge 0$ and $C = e^{1/2}$, it holds that

$$\Pr\left[|X_t| \ge \Delta\right] \le C\cdot e^{-\epsilon\Delta}.$$

Proof. We prove the lemma by an induction on t. The inequality clearly holds for $t = 0$ and all $\Delta \ge 0$. Now, assuming the inequality for X_{t-1} and all $\Delta \ge 0$, we prove the X_t case. When $\Delta \le 1$, the inequality holds trivially, since we have $\epsilon\Delta \le 1/2$, which implies

$$\Pr\left[|X_t| \ge \Delta\right] \le 1 = C\cdot e^{-1/2} \le C\cdot e^{-\epsilon\Delta}.$$

It remains to handle the $\Delta > 1$ case. In order to reach $|X_t| \ge \Delta > 1$, we must have $X_{t-1} \ne 0$; otherwise we would have $X_t \in \{-1/2, 1/2\}$. Furthermore, one of the following two must hold:

- $|X_{t-1}| \ge \Delta - (1/2 - \epsilon)$ and $\mathrm{sgn}(x_t - 1/2) = \mathrm{sgn}(X_{t-1})$.
- $|X_{t-1}| \ge \Delta + (1/2 + \epsilon)$ and $\mathrm{sgn}(x_t - 1/2) = -\mathrm{sgn}(X_{t-1})$.

Note that by the inductive hypothesis, $|X_{t-1}| \geq \Delta - (1/2 - \epsilon)$ holds with probability at most $C \cdot e^{-\epsilon[\Delta-(1/2-\epsilon)]}$. In addition, conditioning on this event, the probability of $\mathrm{sgn}(x_t - 1/2) = \mathrm{sgn}(X_{t-1})$ is still $1/2$ by independence. Thus, the probability of the former is at most $\frac{C}{2}e^{-\epsilon\Delta+\epsilon/2-\epsilon^2}$. An analogous argument upper bounds the probability of the latter condition by $\frac{C}{2}e^{-\epsilon\Delta-\epsilon/2-\epsilon^2}$.

To conclude the inductive step, we need the inequality

$$\frac{C}{2}e^{-\epsilon\Delta+\epsilon/2-\epsilon^2} + \frac{C}{2}e^{-\epsilon\Delta-\epsilon/2-\epsilon^2} \leq C \cdot e^{-\epsilon\Delta},$$

which is equivalent to

$$e^{\epsilon/2} + e^{-\epsilon/2} \leq 2e^{\epsilon^2}.$$

The last inequality can be shown to hold for all $\epsilon \geq 0$ via Taylor expansion. This completes the proof. □

Lemma 11. *On a sequence of T independent random bits, the smooth calibration error incurred by Algorithm 1 with $\epsilon = T^{-1/3}$ is $O(T^{1/3})$ in expectation.*

Proof. Note that Algorithm 1 only predicts three different values: $1/2$, $1/2 + \epsilon$, and $1/2 - \epsilon$. In light of Lemma 9, it suffices to upper bound the expectation of the following three terms at time T: (1) $|\Delta_{1/2} + \Delta_{1/2+\epsilon} + \Delta_{1/2-\epsilon}|$; (2) $|\Delta_{1/2+\epsilon}|$; (3) $|\Delta_{1/2-\epsilon}|$.

The first term. The first part is done by Lemma 10: The stochastic process $S_t \coloneqq \sum_{t'=1}^{t}(x_{t'} - p_{t'})$ exactly matches the one defined in Lemma 10. Therefore, the first term is exactly the absolute value of $S_T = \sum_{t=1}^{T}(x_t - p_t)$. By Lemma 10, we have

$$\mathbb{E}\left[|S_T|\right] = \int_0^{+\infty} \Pr\left[|S_T| \geq \tau\right] \, \mathrm{d}\tau \leq e^{1/2} \int_0^{+\infty} e^{-\epsilon\tau} \, \mathrm{d}\tau = \frac{e^{1/2}}{\epsilon} = O(T^{1/3}).$$

The second term. To analyze the second part, it is convenient to assume that the nature samples random bits $b_1, b_2, \ldots, b_T \sim \mathsf{Bernoulli}(1/2)$ independently at the beginning, and uses these bits one by one as the outcomes for the steps on which $1/2 + \epsilon$ is predicted. More formally, whenever the forecaster predicts $1/2 + \epsilon$ at time t, the nature calculates $k = \sum_{t'=1}^{t} \mathbb{1}\left[p_{t'} = 1/2 + \epsilon\right]$ and sets $x_t = b_k$. Note that this change does not alter the distribution of the random outcomes, and thus the execution of Algorithm 1 remains unchanged.

Then, we note that $\Delta_{1/2+\epsilon}$ can be written as

$$\sum_{i=1}^{m} b_i - m \cdot (1/2 + \epsilon),$$

where m is the number of times $1/2 + \epsilon$ is predicted. By a Chernoff bound and a union bound over $m \in [T]$, for any $\delta \in (0, 1)$, it holds with probability $1 - \delta$ that for all $m \in [T]$,

$$\left|\sum_{i=1}^{m} b_i - \frac{m}{2}\right| \leq \sqrt{\frac{T \ln(2T/\delta)}{2}}.$$

The above implies

$$|\Delta_{1/2+\epsilon}| \leq \left|\sum_{i=1}^{m} b_i - \frac{m}{2}\right| + m\epsilon \leq \sqrt{\frac{T \ln(2T/\delta)}{2}} + T\epsilon.$$

Setting $\delta = 1/T$ and $\epsilon = 1/T^{1/3}$ shows that $|\Delta_{1/2+\epsilon}| = O(T^{2/3})$ with probability $1 - 1/T$. Finally, since $|\Delta_{1/2+\epsilon}|$ is always upper bounded by T, we have

$$\mathbb{E}\left[|\Delta_{1/2+\epsilon}|\right] = O(T^{2/3}) + \delta \cdot T = O(T^{2/3}).$$

Wrapping up. By an analogous argument to the above, we have $\mathbb{E}\left[|\Delta_{1/2-\epsilon}|\right] = O(T^{2/3})$. Finally, by Lemma 9, the expected smooth calibration error is upper bounded by

$$\mathbb{E}\left[|S_T|\right] + \epsilon\,\mathbb{E}\left[|\Delta_{1/2+\epsilon}|\right] + \epsilon\,\mathbb{E}\left[|\Delta_{1/2-\epsilon}|\right] \leq O(T^{1/3}) + 2T^{-1/3} \cdot O(T^{2/3}) = O(T^{1/3}).$$

□

Corollary 12. *The calibration distance incurred by Algorithm 1 with $\epsilon = T^{-1/3}$ is $O(T^{1/3})$ in expectation.*

Proof. Lemmas 3 and 11 imply that Algorithm 1 incurs an $O(T^{1/3})$ lower calibration distance in expectation. Since Algorithm 1 predicts at most three different values, the corollary follows from the second part of Theorem 2. □

7.2 A Polylogarithmic Calibration Distance for Random Bits

In the previous section, we saw how a simple strategy improves the calibration distance from $\Theta(T^{1/2})$ to $O(T^{1/3})$ on a random bit sequence of length T. It turns out that applying the same idea in a slightly more involved way would reduce the distance significantly, to $O(\log^{3/2} T)$.

The forecaster's strategy starts by predicting the value $1/2$ for $T/2$ steps. After that, we expect an $O(\sqrt{T})$ gap between the counts of ones and zeros so far. Say that the number of ones is larger. Then, in the remaining $T/2$ steps, the forecaster keeps predicting $1/2 + \sqrt{\ln T/T}$, until the sum of x_t's and the sum of p_t's are roughly the same. The key observation is that, when the forecaster succeeds in bringing this difference down to zero, the expected smooth calibration error so far is merely $O(\sqrt{\log T})$. For the remaining time steps, we recursively apply the same strategy (with a smaller value of T), and the process must end in $O(\log T)$ rounds. It is relatively easy to show that the errors in different rounds can be aggregated together to give the $O(\log^{3/2} T)$ upper bound.

Formally, we state the strategy of the forecaster in Algorithm 2.

We will prove the following bound, which immediately implies Proposition 1.

Lemma 13. *On a sequence of T independent random bits, the smooth calibration error incurred by Algorithm 2 is $O(\log^{3/2} T)$ in expectation.*

Proof of Proposition 1. Note that in Algorithm 2, the outer for-loop is executed at most $O(\log T)$ times. Furthermore, each prediction made by the forecaster is either $1/2$, or a value $p_i^{(r)}$ specific to a round r. Therefore, $p_1, p_2, \ldots, p_T$ contain at most $O(\log T)$ different values. Then, by Lemma 3, Lemma 13 and the second part of Theorem 2, Algorithm 2 achieves

$$\mathbb{E}\left[\mathsf{CalDist}(x, p)\right] \leq O(1) \cdot \mathbb{E}\left[\mathsf{smCE}(x, p)\right] + O(\log T) = O(\log^{3/2} T).$$

□

Algorithm 2: Adaptive-Bias Forecaster for Random Bits

Input: Time horizon T. Parameter $\epsilon > 0$. Online access to $x_1, x_2, \ldots, x_T$.

Output: Predictions $p_1, p_2, \ldots, p_T$.

$t \leftarrow 0$;
for $r = 1, 2, 3, \ldots$ **do**
  $T^{(r)} \leftarrow T - t$;
  for $i = 1, 2, \ldots, T^{(r)}/2$ **do**
    $t \leftarrow t + 1$;
    $p_t \leftarrow 1/2$; $p_i^{(r)} \leftarrow 1/2$; Predict p_t;
    Observe x_t; $x_i^{(r)} \leftarrow x_t$;
  end
  $\Delta^{(r)} \leftarrow \sum_{i=1}^{T^{(r)}/2} (x_i^{(r)} - p_i^{(r)})$;
  $\epsilon^{(r)} \leftarrow \operatorname{sgn}\left(\Delta^{(r)}\right) \cdot \min\left\{ \frac{2\left|\Delta^{(r)}\right|}{T^{(r)}} + \sqrt{\frac{\ln T^{(r)}}{T^{(r)}}}, \frac{1}{2} \right\}$;
  for $i = T^{(r)}/2 + 1, \ldots, T^{(r)}$ **do**
    $t \leftarrow t + 1$;
    $p_t \leftarrow 1/2 + \epsilon^{(r)}$; $p_i^{(r)} \leftarrow 1/2 + \epsilon^{(r)}$; Predict p_t;
    Observe x_t; $x_i^{(r)} \leftarrow x_t$;
    if $\sum_{j=1}^{i} \left(x_j^{(r)} - p_j^{(r)} \right) \in [-1, 1]$ **then break**;
  end
  if $t = T$ **then break**;
end

Our proof Lemma 13 is decomposed into two parts: First, we argue that it suffices to bound the expected smooth calibration error on outcomes $x^{(r)}$ and predictions $p^{(r)}$, and their sum gives an upper bound on $\mathsf{smCE}(x, p)$. Second, we show that for every r, the error is bounded by $O(\sqrt{\log T})$ in expectation. Lemma 13 then directly follows, since there are at most $O(\log T)$ rounds.

Formally, we have the following two lemmas. The first lemma simply states that the smooth calibration error is subadditive with respect to sequence concatenation.

Lemma 14. *Let $x^{(1)} \in \{0,1\}^{\mathsf{len}^{(1)}}$, ..., $x^{(R)} \in \{0,1\}^{\mathsf{len}^{(R)}}$ be binary sequences, and $p^{(1)} \in [0,1]^{\mathsf{len}^{(1)}}$, ..., $p^{(R)} \in [0,1]^{\mathsf{len}^{(R)}}$ be sequences with the corresponding lengths. Let x and p be the concatenations of $x^{(r)}$ and $p^{(r)}$ in ascending order. Then,*

$$\mathsf{smCE}(x, p) \le \sum_{r=1}^{R} \mathsf{smCE}(x^{(r)}, p^{(r)}).$$

The second lemma bounds the expected smooth calibration error in each round.

Lemma 15. *Let $R \coloneqq \lceil \log_2 T \rceil$. Over the randomness in the execution of Algorithm 2, define random variables $X_1, X_2, \ldots, X_R$ as follows: For every $r \in [R]$, if the algorithm reaches the r-th round, $X_r = \mathsf{smCE}(x^{(r)}, p^{(r)})$; otherwise, $X_r = 0$. Then, for every $r \in [R]$,*

$$\mathbb{E}[X_r] = O(\sqrt{\log T}),$$

where $O(\cdot)$ hides a universal constant that is independent of T and r.

It is clear that Lemma 13 is a directly corollary of Lemmas 14 and 15.

Proof of Lemma 14. By definition of the smooth calibration error, we have

$$\begin{aligned}
\mathsf{smCE}(x, p) &= \sup_{f \in \mathcal{F}} \left[\sum_{r=1}^{R} \sum_{t=1}^{\mathsf{len}^{(r)}} f(p_t^{(r)}) \cdot (x_t^{(r)} - p_t^{(r)}) \right] \\
&\le \sum_{r=1}^{R} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{\mathsf{len}^{(r)}} f(p_t^{(r)}) \cdot (x_t^{(r)} - p_t^{(r)}) \right] \\
&= \sum_{r=1}^{R} \mathsf{smCE}(x^{(r)}, p^{(r)}).
\end{aligned}$$

□

Proof of Lemma 15. We fix a round $r \in [R]$, and condition on the event that $T^{(r)} = L$ holds at the beginning of the r-th iteration of the outer for-loop in Algorithm 2. Note that the event $T^{(r)} = L$ only depends on the first $T - L$ bits $x_1, x_2, \ldots, x_{T-L}$, so conditioning on $T^{(r)} = L$ does not change the distribution of the remaining random bits x_{T-L+1} through x_T.

We say that Round r of Algorithm 2 *succeeds* if the round ends by taking the break on Line 15 in the inner for-loop; Round r *fails* otherwise.

The case that Round r succeeds. We start by upper bounding the value of X_r assuming that the r-th round succeeds. Let $\mathsf{len}^{(r)}$ denote the number of steps in Round r. For $\alpha \in \{1/2, 1/2+\epsilon^{(r)}\}$, we define Δ_α as the total bias incurred by the steps on which α is predicted during Round r, i.e.,

$$\Delta_\alpha \coloneqq \sum_{i=1}^{\mathsf{len}^{(r)}} \left(x_i^{(r)} - p_i^{(r)}\right) \cdot \mathbb{1}\left[p_i^{(r)} = \alpha\right].$$

By definition of the smooth calibration error, we have

$$\begin{aligned}
X_r &= \mathsf{smCE}\left(x^{(r)}, p^{(r)}\right) = \sup_{f\in\mathcal{F}} \left[f(1/2)\Delta_{1/2} + f(1/2+\epsilon^{(r)})\Delta_{1/2+\epsilon^{(r)}}\right] \\
&= \sup_{f\in\mathcal{F}} \left[f(1/2)(\Delta_{1/2} + \Delta_{1/2+\epsilon^{(r)}}) + (f(1/2+\epsilon^{(r)}) - f(1/2))\Delta_{1/2+\epsilon^{(r)}}\right] \\
&\le \sup_{f\in\mathcal{F}} \left[f(1/2)(\Delta_{1/2} + \Delta_{1/2+\epsilon^{(r)}})\right] + \sup_{f\in\mathcal{F}} \left[(f(1/2+\epsilon^{(r)}) - f(1/2))\Delta_{1/2+\epsilon^{(r)}}\right] \\
&\le \left|\Delta_{1/2} + \Delta_{1/2+\epsilon^{(r)}}\right| + \left|\epsilon^{(r)}\right| \cdot \left|\Delta_{1/2+\epsilon^{(r)}}\right|,
\end{aligned}$$

where the last step follows since every $f \in \mathcal{F}$ is 1-Lipschitz and bounded between -1 and $+1$.

When Round r succeeds, we have

$$\Delta_{1/2} + \Delta_{1/2+\epsilon^{(r)}} = \sum_{i=1}^{\mathsf{len}^{(r)}} \left(x_i^{(r)} - p_i^{(r)}\right) \in [-1, 1],$$

which further implies $\left|\Delta_{1/2+\epsilon^{(r)}}\right| \le |\Delta_{1/2}| + 1$.

Therefore, assuming the success of Round r, we have

$$X_r \le 1 + \left|\epsilon^{(r)}\right| \cdot \left(\left|\Delta_{1/2}\right| + 1\right) \le \frac{3}{2} + \frac{2\Delta_{1/2}^2}{L} + |\Delta_{1/2}|\sqrt{\frac{\ln L}{L}}.$$

Control the failure probability. To show that Round r succeeds with high probability, we first argue that when Algorithm 2 chooses $\epsilon^{(r)}$ in Line 10, the minimum takes the first value with high probability. Indeed, this is true as long as

$$\frac{2|\Delta^{(r)}|}{L} + \sqrt{\frac{\ln L}{L}} \le \frac{1}{2},$$

which is equivalent to

$$|\Delta^{(r)}| \le \left(\frac{1}{4} - \frac{1}{2}\sqrt{\frac{\ln L}{L}}\right) \cdot L.$$

For sufficiently large L, we have $\frac{1}{2}\sqrt{\ln L/L} \le 1/8$, so the above is true as long as $|\Delta^{(r)}| \le L/8$. Noting that $\Delta^{(r)}$ is the difference between a sample from $\mathsf{Binomial}(L/2, 1/2)$ and its mean $L/4$, it follows from a Chernoff bound that $|\Delta^{(r)}| \le L/8$ holds with probability $1 - e^{-\Omega(L)}$.

Now, assuming that $\epsilon^{(r)}$ satisfies the aforementioned condition, what is the probability for Round r to fail? Without loss of generality, assume that $\Delta^{(r)} \ge 0$. Then, the failure of Round r would imply $\Delta^{(r)} + \sum_{i=L/2+1}^{L}(x_i^{(r)} - p_i^{(r)}) \ge 0$; otherwise there must be a time step $i \in [L/2+1, L]$ at

which $\sum_{j=1}^{i}(x_j^{(r)} - p_j^{(r)})$ falls into the interval $[-1, 1]$, which allows the round to end. Recalling that $p_i^{(r)} = 1/2+\epsilon^{(r)}$ for every $i \in [L/2+1, L]$, we can rewrite the inequality $\Delta^{(r)}+\sum_{i=L/2+1}^{L}(x_i^{(r)}-p_i^{(r)}) \geq 0$ as

$$\sum_{i=L/2+1}^{L} x_i^{(r)} - \frac{L}{4} \geq -\Delta^{(r)} + \frac{L}{2}\epsilon^{(r)} = \frac{1}{2}\sqrt{L \ln L},$$

where the second step applies $\epsilon^{(r)} = 2\Delta^{(r)}/L + \sqrt{\ln L/L}$. Again, since the left-hand side above is a Binomial random variable (from $\mathsf{Binomial}(L/2, 1/2)$) minus its mean, the probability for the above inequality to hold is, by a Chernoff bound, at most

$$\exp\left(-2 \cdot \frac{L}{2} \cdot \left(\frac{\sqrt{L \ln L}/2}{L/2}\right)^2\right) = \frac{1}{L}.$$

Therefore, we conclude that the probability for Round r to fail (conditioning on $T^{(r)} = L$) is at most $e^{-\Omega(L)} + \frac{1}{L} = O(1/L)$.

Put everything together. Our analysis for the case that Round r succeeds, together with the observation that X_r is always at most L, implies that when $T^{(r)} = L$,

$$X_r \leq \frac{3}{2} + \frac{2\Delta_{1/2}^2}{L} + |\Delta_{1/2}|\sqrt{\frac{\ln L}{L}} + L \cdot \mathbb{1}\left[\text{Round } r \text{ fails}\right]$$

always holds. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[X_r | T^{(r)} = L\right] &\leq \mathbb{E}\left[\frac{3}{2} + \frac{2\Delta_{1/2}^2}{L} + |\Delta_{1/2}|\sqrt{\frac{\ln L}{L}}\middle| T^{(r)} = L\right] + L \cdot \Pr\left[\text{Round } r \text{ fails}\middle| T^{(r)} = L\right] \\
&\leq O(\sqrt{\log L}) + L \cdot O(1/L) \\
&= O(\sqrt{\log T}). && (L \leq T)
\end{aligned}$$

The second step above applies the observation that conditioning on $T^{(r)} = L$, $\Delta_{1/2}$ is the difference between a sample from $\mathsf{Binomial}(L/2, 1/2)$ and its mean $L/4$, which implies $\mathbb{E}\left[\Delta_{1/2}^2\right] = O(L)$ and $\mathbb{E}\left[|\Delta_{1/2}|\right] = O(\sqrt{L})$.

Finally, the bound on $\mathbb{E}[X_r]$ follows from taking an expectation over the value of $L = T^{(r)}$. □

8 Proof of the Lower Bound

We prove Theorem 4 in this section. It is sufficient to lower bound the smooth calibration error incurred by the forecaster, since by Remark 2 and Lemma 3, we have

$$\mathsf{CalDist}(x, p) \geq \mathsf{LowerCalDist}(x, p) \geq \frac{1}{2}\mathsf{smCE}(x, p).$$

Recall from Lemma 13 that, on a random bit sequence, a forecaster *might* achieve an $o(T^{1/3})$ smooth calibration error in the end. The following lemma states that, in this case, the forecaster must incur an $\Omega(T^{1/3})$ bias—defined as the difference between the total outcome and the total predictions—at some point. Later, we prove Theorem 4 by giving a simple adaptive strategy for the adversary that aims to catch this large bias.

Lemma 16. *There exists a universal constant $c > 0$ such that the following holds for all sufficiently large T and every forecaster $\mathcal{A}$: For $t \in [T]$, let random variable S_t denote the value of $\sum_{t'=1}^{t}(x_{t'} - p_{t'})$ when $\mathcal{A}$ is executed against T independent random bits. Then, at least one of the following two holds:*

$$\Pr\left[\max_{t\in[T]} |S_t| \geq cT^{1/3}\right] \geq c$$

or

$$\mathbb{E}\left[\mathsf{smCE}(x,p)\right] \geq cT^{1/3},$$

where the probability and expectation are over the randomness in both the random bits and the algorithm $\mathcal{A}$.

We first show how Theorem 4 follows from Lemma 16. Given an algorithm for the forecaster, if the second condition in the lemma holds, we immediately get the desired lower bound. If the first condition holds, we let the adversary keep outputting independent random bits until $|S_t|$ reaches $cT^{1/3}$, at which point the adversary deviates from giving random bits, and starts outputting a fixed bit instead. The key is to ensure that after the adversary deviates, the smooth calibration error at the end of the T steps is still $\Omega(T^{1/3})$.

Proof of Theorem 4. Let c be the constant in Lemma 16. Consider the following *mixed* adversary:

- First, the adversary decides whether it is *oblivious* or *adaptive* uniformly at random.
- If the adversary decides to be oblivious, output T independent random bits; otherwise, proceed with the following steps.
- At each step t, independently draw the outcome $x_t \sim \mathsf{Bernoulli}(1/2)$, until the game ends or $S_t \geq cT^{1/3}$ holds at some point.
- If $S_t > 0$, keep outputting bit 1 for the remaining $T - t$ steps; otherwise, output zeros for the rest of the game.

Fix an arbitrary algorithm $\mathcal{A}$. Since $\mathsf{CalDist}(x,p) \geq \frac{1}{2}\mathsf{smCE}(x,p)$, it suffices to prove that $\mathcal{A}$ incurs an $\Omega(T^{1/3})$ smooth calibration error against the mixed adversary defined above. By Lemma 16, at least one of the two conditions in the lemma must hold when $\mathcal{A}$ runs on T random bits. If the latter holds, i.e., $\mathbb{E}\left[\mathsf{smCE}(x,p)\right] \geq cT^{1/3}$, we get the desired lower bound. This is because the mixed adversary chooses to be oblivious with probability $1/2$, which implies a lower bound of $cT^{1/3}/2 = \Omega(T^{1/3})$ on the expected smooth calibration error.

Otherwise, assume that the former condition holds. Consider two instances of the algorithm, denoted by $\mathcal{A}_1$ and $\mathcal{A}_2$, such that $\mathcal{A}_1$ runs on the oblivious adversary, whereas $\mathcal{A}_2$ runs against the adaptive adversary. Importantly, the two instances are coupled such that the two adversaries use the same random bits, and $\mathcal{A}_1$ and $\mathcal{A}_2$ share their internal randomness.

Let $\mathcal{E}$ denote the event that, for instance $\mathcal{A}_1$, $\max_{t\in[T]} |S_t| \geq cT^{1/3}$ holds. We will show that, whenever event $\mathcal{E}$ happens, the other instance $\mathcal{A}_2$ gives $\mathsf{smCE}(x,p) \geq cT^{1/3}$. It would then follow from the first condition in Lemma 16 that

$$\Pr_{\mathcal{A}_2}\left[\mathsf{smCE}(x,p) \geq cT^{1/3}\right] \geq \Pr_{\mathcal{A}_1}\left[\mathcal{E}\right] \geq c.$$

Finally, since the mixed adversary decides to be adaptive with probability $1/2$, we have a lower bound of $\frac{c}{2} \cdot cT^{1/3} = \Omega(T^{1/3})$ on the smooth calibration error when $\mathcal{A}$ faces the mixed adversary.

Now assume that $\mathcal{E}$ happens. Let t be the first time step at which $|S_t| \geq cT^{1/3}$ holds. By definition, the adaptive adversary in the $\mathcal{A}_2$ instance deviates from the random bits at time t. If $S_t > 0$, we have $x_{t+1} = x_{t+2} = \cdots = x_T = 1$, which implies

$$\mathsf{smCE}(x,p) \geq \sum_{t'=1}^{T}(x_{t'} - p_{t'}) \geq \sum_{t'=1}^{t}(x_{t'} - p_{t'}) = S_t \geq cT^{1/3}.$$

Similarly, if $S_t < 0$, we get $x_{t+1} = \cdots = x_T = 0$, which gives

$$\mathsf{smCE}(x,p) \geq \sum_{t'=1}^{T}(-1)\cdot(x_{t'} - p_{t'}) = \sum_{t'=1}^{T}(p_{t'} - x_{t'}) \geq \sum_{t'=1}^{t}(p_{t'} - x_{t'}) = -S_t \geq cT^{1/3}.$$

This completes the proof. □

To prove Lemma 16, we use the following standard anti-concentration bound for Binomial distributions, for which we give a proof in Appendix C for completeness.

Lemma 17. *For all sufficiently large integer n,*

$$\Pr_{X\sim\mathsf{Binomial}(n,1/2)}\left[|X - n/2| \geq \sqrt{n}/10\right] \geq \frac{3}{4}.$$

Now, we prove Lemma 16.

Proof of Lemma 16. We prove the lemma for $c = 1/3600$. Fix an algorithm $\mathcal{A}$. Let random variable ϵ_t denote the value of $p_t - 1/2$ when we run $\mathcal{A}$ on a sequence of T random bits. The key quantity in our proof will be $X \coloneqq \sum_{t=1}^{T}\epsilon_t^2$.

Case 1. X is large in expectation. If $\mathbb{E}[X] \geq cT^{1/3}$, we argue that $\mathsf{smCE}(x,p)$ will be large as well. Consider the function $f : v \mapsto 1/2 - v$ over $[0,1]$, which is clearly 1-Lipschitz and bounded between -1 and 1. By definition of the smooth calibration error,

$$\mathsf{smCE}(x,p) \geq \sum_{t=1}^{T} f(p_t)\cdot(x_t - p_t) = \sum_{t=1}^{T}(1/2 - p_t)\cdot(x_t - p_t).$$

For every $t \in [T]$, note that x_t and p_t are independent conditioning on $x_{1:(t-1)}$ and $p_{1:(t-1)}$. Thus, we have $\mathbb{E}[(1/2 - p_t)\cdot(x_t - 1/2)] = 0$, which implies

$$\mathbb{E}[(1/2 - p_t)\cdot(x_t - p_t)] = \mathbb{E}[(1/2 - p_t)\cdot(x_t - 1/2)] - \mathbb{E}[(1/2 - p_t)\cdot(p_t - 1/2)] = \mathbb{E}\left[\epsilon_t^2\right].$$

We conclude that $\mathbb{E}[\mathsf{smCE}(x,p)] \geq \mathbb{E}\left[\sum_{t=1}^{T}\epsilon_t^2\right] = \mathbb{E}[X] \geq cT^{1/3}$.

Case 2. X is small in expectation. Pick $L = \lfloor T^{2/3} \rfloor$ and let $m = \lfloor T/L \rfloor \geq T^{1/3}$. We divide the first $mL \leq T$ steps into m epochs, each of length L. For $i \in [m]$, let $\mathcal{T}_i \coloneqq \{(i-1) \cdot L + 1, (i-1) \cdot L + 2, \ldots, i \cdot L\}$ denote the time steps in the i-th epoch. We say that the i-th epoch is *good*, if

$$\left| \sum_{t \in \mathcal{T}_i} x_t - \frac{L}{2} \right| \geq \frac{\sqrt{L}}{10}.$$

The i-th epoch is called *weak*, if

$$\sum_{t \in \mathcal{T}_i} \epsilon_t^2 \leq \frac{1}{400}.$$

We first note that with high probability, there are many good epochs and many weak epochs. For good epochs, the claim follows from Lemma 17 and a Chernoff bound: The quantity $\sum_{t \in \mathcal{T}_i} x_t$ follows the Binomial distribution $\mathsf{Binomial}(L, 1/2)$. By Lemma 17, the probability for each epoch to be good is at least $3/4$, as long as L is sufficiently large. By a Chernoff bound, with probability at least $1 - \exp(-\Omega(m))$, there are at least $\frac{2}{3}m$ good epochs. Again, when T is sufficiently large, the failure probability $\exp(-\Omega(m))$ is at most $1/3$.

For weak epochs, recall that we are under the assumption that $\mathbb{E}[X] = \mathbb{E}\left[\sum_{t=1}^{T} \epsilon_t^2\right] \leq cT^{1/3} = T^{1/3}/3600$. The expected number of epochs that are not weak is then at most $\frac{T^{1/3}/3600}{1/400} = \frac{1}{9}T^{1/3}$. By Markov's inequality, the probability that there are $\geq \frac{1}{3}T^{1/3}$ epochs that are not weak is at most $\frac{\frac{1}{9}T^{1/3}}{\frac{1}{3}T^{1/3}} = \frac{1}{3}$. In other words, with probability at least $2/3$, there are at least $\frac{2}{3}T^{1/3} \geq \frac{2}{3}m$ weak epochs.

By a union bound, with probability at least $1/3$, there are at least $\frac{2}{3}m$ good epochs and at least $\frac{2}{3}m$ weak epochs. In particular, with probability at least $1/3$, there exists an epoch i that is both good and weak. Recall that S_t is the partial sum of $(x_t - p_t)$. We have

$$S_{iL} - S_{(i-1)L} = \sum_{t \in \mathcal{T}_i} (x_t - p_t) = \sum_{t \in \mathcal{T}_i} (x_t - 1/2) - \sum_{t \in \mathcal{T}_i} (p_t - 1/2) = \left(\sum_{t \in \mathcal{T}_i} x_t - \frac{L}{2} \right) - \sum_{t \in \mathcal{T}_i} \epsilon_t.$$

By definition of good and weak epochs, we have $\left|\sum_{t \in \mathcal{T}_i} x_t - \frac{L}{2}\right| \geq \frac{\sqrt{L}}{10}$ and $\left|\sum_{t \in \mathcal{T}_i} \epsilon_t\right| \leq \sqrt{L} \cdot \sqrt{\sum_{t \in \mathcal{T}_i} \epsilon_t^2} \leq \frac{\sqrt{L}}{20}$, where the second bound follows from the Cauchy-Schwarz inequality. These two bounds further imply

$$|S_{iL} - S_{(i-1)L}| \geq \frac{\sqrt{L}}{10} - \frac{\sqrt{L}}{20} = \frac{\sqrt{L}}{20},$$

and thus,

$$\max_{t \in [T]} |S_t| \geq \max\{S_{iL}, S_{(i-1)L}\} \geq \frac{\sqrt{L}}{40}.$$

Recall that $L = \lfloor T^{2/3} \rfloor$ and $c = 1/3600$. For all sufficiently large T, the above gives a stronger guarantee than $\Pr\left[\max_{t \in [T]} |S_t| \geq cT^{1/3}\right] \geq c$. This completes the proof. □

A Basic Facts about Calibration Measures

We first show that the calibration distance is always upper bounded by the ECE.

Proposition 18. *For any $x \in \{0,1\}^T$ and $p \in [0,1]^T$,*

$$\mathsf{CalDist}(x,p) \le \mathsf{ECE}(x,p).$$

Proof. For each $t \in [T]$, define

$$q_t \coloneqq g(p_t) \coloneqq \frac{\sum_{t'=1}^{T} x_{t'} \cdot \mathbb{1}\left[p_{t'} = p_t\right]}{\sum_{t'=1}^{T} \mathbb{1}\left[p_{t'} = p_t\right]}$$

as the actual frequency of ones when the value p_t is predicted. Note that q_t is completely determined by p_t.

We first show that q is in $\mathcal{C}(x)$: For any $\alpha \in [0,1]$, it holds that

$$\begin{aligned}
\sum_{t=1}^{T}(x_t - q_t)\cdot \mathbb{1}\left[q_t = \alpha\right] &= \sum_{\beta\in[0,1]} \mathbb{1}\left[g(\beta)=\alpha\right] \sum_{t=1}^{T}(x_t-\alpha)\cdot\mathbb{1}\left[p_t=\beta\right] \\
&= \sum_{\beta\in[0,1]} \mathbb{1}\left[g(\beta)=\alpha\right]\left[\sum_{t=1}^{T} x_t\cdot\mathbb{1}\left[p_t=\beta\right] - g(\beta)\cdot\sum_{t=1}^{T}\mathbb{1}\left[p_t=\beta\right]\right] \\
&= 0,
\end{aligned}$$

where the last step follows from the definition of $g(\cdot)$. This shows $q \in \mathcal{C}(x)$.

Then, we compute the distance $\|p-q\|_1$:

$$\|p-q\|_1 = \sum_{t=1}^{T}|p_t - q_t| = \sum_{\beta\in[0,1]}\sum_{t=1}^{T}|p_t-q_t|\cdot\mathbb{1}\left[p_t=\beta\right].$$

Fix $\beta \in \{p_1, p_2, \ldots, p_T\}$. We note that $p_t = \beta$ implies $q_t = g(\beta)$. It follows that

$$\begin{aligned}
\sum_{t=1}^{T}|p_t-q_t|\cdot\mathbb{1}\left[p_t=\beta\right] &= |\beta - g(\beta)|\cdot\sum_{t=1}^{T}\mathbb{1}\left[p_t=\beta\right] \\
&= \left|\beta\cdot\sum_{t=1}^{T}\mathbb{1}\left[p_t=\beta\right] - g(\beta)\cdot\sum_{t=1}^{T}\mathbb{1}\left[p_t=\beta\right]\right| \\
&= \left|\sum_{t=1}^{T} p_t\cdot\mathbb{1}\left[p_t=\beta\right] - \sum_{t=1}^{T} x_t\cdot\mathbb{1}\left[p_t=\beta\right]\right| \\
&= \left|\sum_{t=1}^{T}(x_t-p_t)\cdot\mathbb{1}\left[p_t=\beta\right]\right|.
\end{aligned}$$

Summing over $\beta \in [0,1]$ gives

$$\|p-q\|_1 = \sum_{\beta\in[0,1]}\left|\sum_{t=1}^{T}(x_t-p_t)\cdot\mathbb{1}\left[p_t=\beta\right]\right| = \mathsf{ECE}(x,p).$$

Since we showed that $q \in \mathcal{C}(x)$ and $\|p - q\|_1 = \mathsf{ECE}(x, p)$, the definition of $\mathsf{CalDist}(x, p)$ immediately implies $\mathsf{CalDist}(x, p) \leq \mathsf{ECE}(x, p)$. □

The following lemma states the Lipschitz continuity of CalDist, LowerCalDist and smCE.

Lemma 19. *The following inequalities hold for any $x \in \{0, 1\}^T$ and $p, \tilde{p} \in [0, 1]^T$:*

$$|\mathsf{CalDist}(x, p) - \mathsf{CalDist}(x, \tilde{p})| \leq \|p - \tilde{p}\|_1,$$

$$|\mathsf{LowerCalDist}(x, p) - \mathsf{LowerCalDist}(x, \tilde{p})| \leq \|p - \tilde{p}\|_1,$$

$$|\mathsf{smCE}(x, p) - \mathsf{smCE}(x, \tilde{p})| \leq 2\|p - \tilde{p}\|_1.$$

Proof. By definition,

$$\mathsf{CalDist}(x, p) = \min_{q \in \mathcal{C}(x)} \|p - q\|_1 \leq \min_{q \in \mathcal{C}(x)} (\|p - \tilde{p}\|_1 + \|\tilde{p} - q\|_1) = \mathsf{CalDist}(x, \tilde{p}) + \|p - \tilde{p}\|_1.$$

By symmetry, we have $\mathsf{CalDist}(x, \tilde{p}) \leq \mathsf{CalDist}(x, p) + \|p - \tilde{p}\|_1$. This proves the first inequality.

The proof for LowerCalDist is analogous:

$$\begin{aligned}
\mathsf{LowerCalDist}(x, p) &= \inf_{\mathcal{D} \in \underline{\mathcal{C}}(x)} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \mathcal{D}_t} [|p_t - q_t|] \\
&\leq \inf_{\mathcal{D} \in \underline{\mathcal{C}}(x)} \left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \mathcal{D}_t} [|p_t - \tilde{p}_t| + |\tilde{p}_t - q_t|] \right] \\
&= \inf_{\mathcal{D} \in \underline{\mathcal{C}}(x)} \left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t \sim \mathcal{D}_t} [|\tilde{p}_t - q_t|] \right] + \|p - \tilde{p}\|_1 \\
&= \mathsf{LowerCalDist}(x, \tilde{p}) + \|p - \tilde{p}\|_1.
\end{aligned}$$

Again, we also have $\mathsf{LowerCalDist}(x, \tilde{p}) \leq \mathsf{LowerCalDist}(x, p) + \|p - \tilde{p}\|_1$. This proves the second inequality.

Let f be a 1-Lipschitz function from $[0, 1]$ to $[-1, 1]$. Note that for any $p, q \in [0, 1]$ and $x \in \{0, 1\}$, we have

$$\begin{aligned}
& |f(p) \cdot (x - p) - f(q) \cdot (x - q)| \\
\leq\ & |f(p) \cdot (x - p) - f(p) \cdot (x - q)| + |f(p) \cdot (x - q) - f(q) \cdot (x - q)| \\
=\ & |f(p)| \cdot |p - q| + |f(p) - f(q)| \cdot |x - q| \\
\leq\ & 2|p - q|.
\end{aligned}$$

The last step follows from $|f(p)| \leq 1$, $|f(p) - f(q)| \leq |p - q|$, and $|x - q| \leq 1$.

Thus, by definition of the smooth calibration error,

$$\begin{aligned}\mathsf{smCE}(x,p) &= \sup_{f\in\mathcal{F}} \sum_{t=1}^{T} f(p_t)\cdot(x_t-p_t)\\ &\le \sup_{f\in\mathcal{F}} \sum_{t=1}^{T} [f(\tilde{p}_t)\cdot(x_t-\tilde{p}_t)+2|p_t-\tilde{p}_t|]\\ &= \sup_{f\in\mathcal{F}} \sum_{t=1}^{T} f(\tilde{p}_t)\cdot(x_t-\tilde{p}_t)+2\|p-\tilde{p}\|_1\\ &= \mathsf{smCE}(x,\tilde{p})+2\|p-\tilde{p}\|_1.\end{aligned}$$

By symmetry, we also have $\mathsf{smCE}(x,\tilde{p}) \le \mathsf{smCE}(x,p)+2\|p-\tilde{p}\|_1$. This completes the proof. □

B Proofs for Section 5

B.1 Failure of Naïve Consolidation

In the context of Lemma 7, we give a concrete example in which the straightforward way of merging the transportation fails to keep the cost low. This explains why the proof of Lemma 7 involves a complicated consolidation strategy.

Let $T=2k$. The outcomes $x=(1,0,\dots,0,1,1,\dots,1,0)$ consist of 1 followed by $k-1$ copies of zeros, $k-1$ copies of ones and a single zero. The predictions $p=(\epsilon,\epsilon,\dots,\epsilon,1-\epsilon,1-\epsilon,\dots,1-\epsilon)$ contain k copies of ϵ followed by k copies of $1-\epsilon$, where $\epsilon=1/(2k)$. Let $\mathcal{D}_1,\dots,\mathcal{D}_k$ be degenerate distributions over $\{1/k\}$ and $\mathcal{D}_{k+1},\dots,\mathcal{D}_{2k}$ be degenerate distributions over $\{1-1/k\}$. Clearly, we have $\mathcal{D}\in\underline{\mathcal{C}}(x)$. Furthermore, the cost of $\mathcal{D}$ is given by

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_t\sim\mathcal{D}_t}[|p_t-q_t|] = T\cdot\frac{1}{2k}=1.$$

With the notation in Section 5.4, we have $s_1=\epsilon$, $s_2=1-\epsilon$, and all the transportation (specified by $\mathcal{D}$) are into the interval $[s_1,s_2]$. However, if we consolidate all these transportation, we would end up with a new destination of $1/2$, and the cost would surge to $\Omega(T)$.

B.2 Technical Lemmas

We state and prove all the technical lemmas used in Section 5 here.

Lemma 20. *For any $x,y\in[0,1]$ and $\Delta_x,\Delta_y\in\mathbb{R}$, there exists a 1-Lipschitz function $f:[0,1]\to[-1,1]$ such that*

$$f(x)\cdot\Delta_x+f(y)\cdot\Delta_y=\begin{cases}|\Delta_x|+|\Delta_y|, & \Delta_x\Delta_y\ge 0,\\ |\Delta_x+\Delta_y|+|x-y|\cdot\min\{|\Delta_x|,|\Delta_y|\}, & \Delta_x\Delta_y<0.\end{cases}$$

Proof. First, suppose that $\Delta_x\Delta_y\ge 0$, in which case there exists $s\in\{\pm 1\}$ such that $s\cdot\Delta_x=|\Delta_x|$ and $s\cdot\Delta_y=|\Delta_y|$. Then, for the constant function $f(v)\coloneqq s$, we have

$$f(x)\cdot\Delta_x+f(y)\cdot\Delta_y=|\Delta_x|+|\Delta_y|.$$

Then, suppose that $\Delta_x\Delta_y < 0$ and $|\Delta_x| \geq |\Delta_y|$. We take $f(v) \coloneqq \operatorname{sgn}(\Delta_x) \cdot (1 - |v - x|)$, which is clearly 1-Lipschiz and bounded between -1 and $+1$. This gives

$$\begin{aligned} f(x) \cdot \Delta_x + f(y) \cdot \Delta_y &= \operatorname{sgn}(\Delta_x) \cdot \Delta_x + \operatorname{sgn}(\Delta_x) \cdot (1 - |x - y|) \cdot \Delta_y \\ &= |\Delta_x| - (1 - |x - y|) \cdot |\Delta_y| \\ &= (|\Delta_x| - |\Delta_y|) + |x - y| \cdot |\Delta_y| \\ &= |\Delta_x + \Delta_y| + |x - y| \cdot \min\{|\Delta_x|, |\Delta_y|\}. \end{aligned}$$

Similarly, when $\Delta_x\Delta_y < 0$ and $|\Delta_x| < |\Delta_y|$, taking $f(v) \coloneqq \operatorname{sgn}(\Delta_y) \cdot (1 - |v - y|)$ also gives

$$\begin{aligned} f(x) \cdot \Delta_x + f(y) \cdot \Delta_y &= \operatorname{sgn}(\Delta_y) \cdot \Delta_y + \operatorname{sgn}(\Delta_y) \cdot (1 - |x - y|) \cdot \Delta_x \\ &= |\Delta_y| - (1 - |x - y|) \cdot |\Delta_x| \\ &= (|\Delta_y| - |\Delta_x|) + |x - y| \cdot |\Delta_x| \\ &= |\Delta_x + \Delta_y| + |x - y| \cdot \min\{|\Delta_x|, |\Delta_y|\}. \end{aligned}$$

□

Lemma 21. *For any $0 \leq \alpha \leq \beta \leq 1$ and $p \in [0, 1]$, we have*

$$p \cdot |p - \alpha| + (1 - p) \cdot |p - \beta| \leq 2 \cdot [|p \cdot (1 - \alpha) - (1 - p) \cdot \beta| + (\beta - \alpha) \cdot \min\{p \cdot (1 - \alpha), (1 - p) \cdot \beta\}] .$$

Proof. We prove the inequality for the following three cases separately.

Case 1: $p \leq \alpha$. In this case, the left-hand side of the inequality gets reduced to

$$p \cdot (\alpha - p) + (1 - p) \cdot (\beta - p) = (1 - p) \cdot \beta - p \cdot (1 - \alpha) \leq |p \cdot (1 - \alpha) - (1 - p) \cdot \beta|,$$

which is clearly upper bounded by the right-hand side.

Case 2: $p \geq \beta$. Similarly, we can simplify the left-hand side to

$$p \cdot (p - \alpha) + (1 - p) \cdot (p - \beta) = p \cdot (1 - \alpha) - (1 - p) \cdot \beta \leq |p \cdot (1 - \alpha) - (1 - p) \cdot \beta|,$$

which is, again, upper bounded by the right-hand side.

Case 3: $p \in (\alpha, \beta)$. In this case, using the identities $x + y = |x - y| + 2\min\{x, y\}$ and

$$p \cdot (1 - \alpha) - p \cdot (p - \alpha) = (1 - p) \cdot \beta - (1 - p) \cdot (\beta - p) = p(1 - p),$$

we can write the left-hand side as:

$$\begin{aligned} &|p \cdot (p - \alpha) - (1 - p) \cdot (\beta - p)| + 2 \cdot \min\{p \cdot (p - \alpha), (1 - p) \cdot (\beta - p)\} \\ = &|p \cdot (1 - \alpha) - (1 - p) \cdot \beta| + 2 \cdot \min\{p \cdot (p - \alpha), (1 - p) \cdot (\beta - p)\}. \end{aligned}$$

Thus, to prove the lemma, it remains to show that:

$$\min\{p \cdot (p - \alpha), (1 - p) \cdot (\beta - p)\} \leq (\beta - \alpha) \cdot \min\{p \cdot (1 - \alpha), (1 - p) \cdot \beta\}. \tag{11}$$

If $p \cdot (p - \alpha) \leq (1 - p) \cdot (\beta - p)$, the minimum on the right-hand side of (11) is also achieved by the first term. Then, it is sufficient to prove that $p \cdot (p - \alpha) \leq (\beta - \alpha) \cdot p \cdot (1 - \alpha)$, which is equivalent to

$$p \leq \beta - \alpha\beta + \alpha^2.$$

Note that the assumption $p \cdot (p - \alpha) \leq (1 - p) \cdot (\beta - p)$ is equivalent to

$$p + p \cdot (\beta - \alpha) \leq \beta.$$

Since $\beta - \alpha \geq 0$ and $p \geq \alpha$, we have

$$\alpha \cdot (\beta - \alpha) \leq p \cdot (\beta - \alpha).$$

Adding the two inequalities above together gives the desired inequality $p \leq \beta - \alpha\beta + \alpha^2$.

The remaining case that $p \cdot (p - \alpha) > (1 - p) \cdot (\beta - p)$ can be dealt with in a similar way. In this case, Inequality (11) is equivalent to

$$\beta - p \leq \beta^2 - \alpha\beta.$$

The assumption $p \cdot (p - \alpha) > (1 - p) \cdot (\beta - p)$ implies

$$\beta - p \leq p \cdot (\beta - \alpha).$$

Applying $\beta - \alpha \geq 0$ and $p \leq \beta$ to the above, we get

$$\beta - p \leq \beta^2 - \alpha\beta$$

as desired. □

Lemma 22. *For any $0 \leq \alpha \leq \beta \leq 1$ and $p \in [0, 1]$, we have*

$$\begin{aligned} &\min\{(1 - p) \cdot |p - \alpha| + p \cdot |p - \beta|, (1 - p) \cdot \alpha + p \cdot (1 - \beta)\} \\ \leq\ &10 \cdot [|(1 - p) \cdot \alpha - p \cdot (1 - \beta)| + (\beta - \alpha) \cdot \min\{(1 - p) \cdot \alpha, p \cdot (1 - \beta)\}]. \end{aligned}$$

Proof. As in the proof of Lemma 21, we consider the following three cases.

Case 1: $p \leq \alpha$. The left-hand side is upper bounded by the first term in the minimum, which, in this case, is given by

$$(1 - p) \cdot (\alpha - p) + p \cdot (\beta - p) = (1 - p) \cdot \alpha - p \cdot (1 - \beta) \leq |(1 - p) \cdot \alpha - p \cdot (1 - \beta)|.$$

Clearly, this is upper bounded by the right-hand side.

Case 2: $p \geq \beta$. Similarly, we can upper bound the left-hand side by

$$(1 - p) \cdot (p - \alpha) + p \cdot (p - \beta) = -(1 - p) \cdot \alpha + p \cdot (1 - \beta) \leq |(1 - p) \cdot \alpha - p \cdot (1 - \beta)|,$$

which, in turn, is at most the right-hand side.

Case 3: $p \in (\alpha, \beta)$**.** We first consider the case that $\beta - \alpha$ is large. Concretely, suppose that $\beta - \alpha \geq 1/5$. If so, by the identity $x + y = |x - y| + 2\min\{x, y\}$, we have

$$\begin{aligned}(1-p)\cdot\alpha + p\cdot(1-\beta) &= |(1-p)\cdot\alpha - p\cdot(1-\beta)| + 2\min\{(1-p)\cdot\alpha, p\cdot(1-\beta)\} \\ &\leq 10|(1-p)\cdot\alpha - p\cdot(1-\beta)| + 10(\beta-\alpha)\min\{(1-p)\cdot\alpha, p\cdot(1-\beta)\},\end{aligned}$$

which implies the desired inequality.

We then focus on the case that $\beta - \alpha < 1/5$. We write $x \coloneqq p - \alpha > 0$ and $y \coloneqq \beta - p > 0$. Note that

$$(1-p)\cdot\alpha - p\cdot(1-\beta) = p\cdot(\beta - p) - (1-p)\cdot(p-\alpha) = py - (1-p)x.$$

We claim that if $(1-p)x$ and py are not close (up to a multiplicative factor), we are done. Formally, suppose that

$$\frac{\min\{(1-p)x, py\}}{\max\{(1-p)x, py\}} \leq \frac{2}{3}.$$

Then, we may upper bound the left-hand side of the desired inequality by

$$(1-p)x + py = \max\{(1-p)x, py\} + \min\{(1-p)x, py\} \leq \frac{5}{3}\max\{(1-p)x, py\}.$$

On the other hand, the right-hand side is lower bounded by its first term, namely,

$$\begin{aligned}10|(1-p)\cdot\alpha - p\cdot(1-\beta)| &= 10|py - (1-p)x| \\ &= 10\max\{(1-p)x, py\} - 10\min\{(1-p)x, py\} \\ &\geq \left(10 - 10\cdot\frac{2}{3}\right)\cdot\max\{(1-p)x, py\} \\ &\geq \frac{5}{3}\cdot\max\{(1-p)x, py\}.\end{aligned}$$

This proves the inequality when $\min\{(1-p)x, py\} \leq \frac{2}{3}\max\{(1-p)x, py\}$ holds.

Finally, we deal with the case that both $\beta - \alpha < 1/5$ and $\min\{(1-p)x, py\} > \frac{2}{3}\max\{(1-p)x, py\}$ hold. Note that the second condition implies $\frac{(1-p)x}{py} > \frac{2}{3}$ and $\frac{py}{(1-p)x} > \frac{2}{3}$. Again, we simplify and relax the desired inequality into

$$\begin{aligned}(1-p)x + py &\leq 10(x+y)\min\{(1-p)\cdot\alpha, p\cdot(1-\beta)\} \\ &= 10(x+y)\cdot[p(1-p) - \max\{(1-p)x, py\}].\end{aligned}$$

We argue that both $(1-p)x$ and py are at most $\frac{3}{10}p(1-p)$. Otherwise, suppose that $(1-p)x > \frac{3}{10}p(1-p)$. This implies $x > \frac{3}{10}p > p/5$. Furthermore, we have

$$py > \frac{2}{3}(1-p)x > \frac{2}{3}\cdot\frac{3}{10}p(1-p),$$

which implies $y > (1-p)/5$. We then obtain $x + y > p/5 + (1-p)/5 = 1/5$, which contradicts $x + y = \beta - \alpha < 1/5$. An analogous argument also rules out the possibility that $py > \frac{3}{10}p(1-p)$.

Therefore, it suffices to prove that

$$(1-p)x + py \leq 10(x+y)\cdot\left[p(1-p) - \frac{3}{10}p(1-p)\right] = 7(x+y)\cdot p(1-p),$$

or, equivalently,

$$(1-p)\cdot\frac{x}{x+y}+p\cdot\frac{y}{x+y}\le 7p(1-p).$$

The first term on the left-hand side above can be upper bounded as follows:

$$(1-p)\cdot\frac{x}{x+y}=(1-p)\cdot\frac{1}{1+y/x}\le(1-p)\cdot\frac{1}{1+\frac{2\cdot(1-p)}{3p}}=\frac{p(1-p)}{\frac{2}{3}+\frac{p}{3}}\le\frac{3}{2}\cdot p(1-p),$$

where the second step applies $y/x>\frac{2\cdot(1-p)}{3p}$, which follows from $\frac{py}{(1-p)x}>\frac{2}{3}$. Similarly, we have

$$p\cdot\frac{y}{x+y}=p\cdot\frac{1}{x/y+1}\le p\cdot\frac{1}{\frac{2p}{3\cdot(1-p)}+1}=\frac{p(1-p)}{1-\frac{1}{3}p}\le\frac{3}{2}\cdot p(1-p).$$

Adding the two inequalities above gives

$$(1-p)\cdot\frac{x}{x+y}+p\cdot\frac{y}{x+y}\le 3p(1-p)\le 7p\cdot(1-p),$$

which implies the desired inequality for the last case, and thus completes the proof. □

C Proof for Section 8

We prove Lemma 17, which is restated below.

Lemma 17. *For all sufficiently large integer n,*

$$\Pr_{X\sim\mathsf{Binomial}(n,1/2)}\left[|X-n/2|\ge\sqrt{n}/10\right]\ge\frac{3}{4}.$$

Proof. The mode of $\mathsf{Binomial}(n,1/2)$ is $\lfloor n/2\rfloor$. When $n=2k$ is even, it holds for every $j\in\{0,1,\ldots,n\}$ that

$$\begin{aligned}\Pr_{X\sim\mathsf{Binomial}(2k,1/2)}[X=j]&\le\Pr_{X\sim\mathsf{Binomial}(2k,1/2)}[X=k]\\&=2^{-2k}\frac{(2k)!}{(k!)^2}\\&=(1+o_n(1))\cdot 2^{-2k}\cdot\frac{\sqrt{2\pi\cdot 2k}\cdot(2k/e)^{2k}}{2\pi k\cdot(k/e)^{2k}}\\&=(1+o_n(1))\cdot\frac{\sqrt{2/\pi}}{\sqrt{n}}.\end{aligned}$$

The third step applies Stirling's approximation $\frac{n!}{\sqrt{2\pi n}(n/e)^n}=1+o_n(1)$. Since $\sqrt{2/\pi}<1$, for

sufficiently large n we have an upper bound of $1/\sqrt{n}$. Similarly, when $n = 2k+1$ is odd, we have

$$\begin{aligned}
\Pr_{X\sim\mathsf{Binomial}(n,1/2)}[X=j] &\le 2^{-(2k+1)}\cdot\frac{(2k+1)!}{k!(k+1)!}\\
&= (1+o_n(1))\cdot 2^{-(2k+1)}\cdot\frac{\sqrt{2\pi(2k+1)}\cdot\left(\frac{2k+1}{e}\right)^{2k+1}}{\sqrt{2\pi k}\cdot\sqrt{2\pi(k+1)}\cdot(k/e)^k\left(\frac{k+1}{e}\right)^{k+1}}\\
&= (1+o_n(1))\cdot\frac{1}{\sqrt{\pi k}}\cdot\left(1+\frac{1}{2k}\right)^k\cdot\left(1-\frac{1}{2(k+1)}\right)^{k+1}\\
&= (1+o_n(1))\cdot\frac{1}{\sqrt{\pi k}}\cdot(e^{1/2}+o_n(1))\cdot(e^{-1/2}+o_n(1))\\
&= (1+o_n(1))\cdot\frac{\sqrt{2/\pi}}{\sqrt{n}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\Pr_{X\sim\mathsf{Binomial}(n,1/2)}\left[|X-n/2|<\sqrt{n}/10\right]\\
&=\sum_{j=0}^{n}\Pr_{X\sim\mathsf{Binomial}(n,1/2)}[X=j]\cdot\mathbb{1}\left[|j-n/2|<\sqrt{n}/10\right]\\
&\le\frac{1}{\sqrt{n}}\cdot\left(2\cdot\frac{\sqrt{n}}{10}+1\right)\le\frac{1}{4},
\end{aligned}$$

where the last step holds for all sufficiently large n. □

References

[BGHN23] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Symposium on Theory of Computing (STOC)*, pages 1727–1740, 2023. 1, 1.1, 1.1, 1.2, 2, 2, 3, 4, 5, 5.1, 5.4

[BGJ+22] Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 29362–29373, 2022. 1.2

[Bri50] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. 1.2

[Daw82] A. P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. 1.2

[Daw85] A. P. Dawid. Calibration-based empirical probability. *The Annals of Statistics*, 13(4):1251–1274, 1985. 1.2

[FH18] Dean P. Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. *Games and Economic Behavior*, 109:271–293, 2018. 1.2, 4

[FH21] Dean P. Foster and Sergiu Hart. Forecast hedging and calibration. *Journal of Political Economy*, 129(12):3447–3490, 2021. 1.2, 4

[FL99] Drew Fudenberg and David K. Levine. An easier way to calibrate. *Games and Economic Behavior*, 29(1-2):131–137, 1999. 1.2

[Fos99] Dean P. Foster. A proof of calibration via blackwell's approachability theorem. *Games and Economic Behavior*, 29(1-2):73–78, 1999. 1.2

[FRST11] Dean P. Foster, Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Complexity-based approach to calibration with checking rules. In *Conference on Learning Theory (COLT)*, pages 293–314, 2011. 1.2

[FV98] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998. 1, 1.1, 1.2

[GJN+22] Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. In *Innovations in Theoretical Computer Science (ITCS)*, pages 82:1–82:24, 2022. 1.2

[GJRR24] Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Symposium on Discrete Algorithms (SODA)*, pages 2725–2792, 2024. 1.2

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017. 1.2

[GR22] Chirag Gupta and Aaditya Ramdas. Faster online calibration without randomization: interval forecasts and the power of two choices. In *Conference on Learning Theory (COLT)*, pages 4283–4309, 2022. 1.2

[Har22] Sergiu Hart. Calibrated forecasts: The minimax proof. *arXiv preprint arXiv:2209.05863*, 2022. 1.1, 1.2, 6

[HJKRR18] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning (ICML)*, pages 1939–1948, 2018. 1.2, 1.2

[KE17] Volodymyr Kuleshov and Stefano Ermon. Estimating uncertainty online against an adversary. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2110–2116, 2017. 1.2

[KF08] Sham M. Kakade and Dean P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008. 1.1, 1.2, 2, 4

[KLST23] Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *Conference on Learning Theory (COLT)*, pages 5143–5145, 2023. 1.2

[KMR17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science (ITCS)*, pages 43:1–43:23, 2017. 1.2

[LNPR22] Daniel Lee, Georgy Noarov, Mallesh Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 29051–29063, 2022. 1.2

[NRRX23] Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional unbiased prediction for sequential decision making. In *OPT 2023: Optimization for Machine Learning*, 2023. 1.2

[OKS23] Princewill Okoroafor, Robert Kleinberg, and Wen Sun. Faster recalibration of an online predictor via approachability. *arXiv preprint arXiv:2310.17002*, 2023. 1.2

[QV21] Mingda Qiao and Gregory Valiant. Stronger calibration lower bounds via sidestepping. In *Symposium on Theory of Computing (STOC)*, pages 456–466, 2021. 1.2, 1.2, 3.4, 4

[RST15a] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research (JMLR)*, 16(1):155–186, 2015. 6.1, 6.2, 5

[RST15b] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161:111–153, 2015. 6