An Elementary Predictor Obtaining $2\sqrt{T}+1$ Distance to Calibration

Eshwar Ram Arunachaleswaran* Natalie Collina† Aaron Roth ‡ Mirah Shi§

Abstract

Błasiok et al. [2023] proposed *distance to calibration* as a natural measure of calibration error that unlike expected calibration error (ECE) is continuous. Recently, Qiao and Zheng [2024] (COLT 2024) gave a non-constructive argument establishing the existence of a randomized online predictor that can obtain $O(\sqrt{T})$ distance to calibration in expectation in the adversarial setting, which is known to be impossible for ECE. They leave as an open problem finding an explicit, efficient, deterministic algorithm. We resolve this problem and give an extremely simple, efficient, deterministic algorithm that obtains distance to calibration error at most $2\sqrt{T}+1$.

*Department of Computer and Information Sciences, University of Pennsylvania.
†Department of Computer and Information Sciences, University of Pennsylvania.
‡Department of Computer and Information Sciences, University of Pennsylvania.
§Department of Computer and Information Sciences, University of Pennsylvania.

1 Introduction

Probabilistic predictions of binary outcomes are said to be *calibrated*, if, informally, they are unbiased conditional on their own predictions. For predictors that are not perfectly calibrated, there are a variety of ways to measure calibration error. Perhaps the most popular measure is Expected Calibration Error (ECE), which measures the average bias of the predictions, weighted by the frequency of the predictions. ECE has a number of difficulties as a measure of calibration, not least of which is that it is discontinuous in the predictions. Motivated by this, Błasiok et al. [2023] propose a different measure: distance to calibration, which measures how far a predictor is in ℓ_1 distance from the nearest perfectly calibrated predictor. In the online adversarial setting, it has been known since Foster and Vohra [1998] how to make predictions with ECE growing at a rate of $O(T^{2/3})$. Qiao and Valiant [2021] show that obtaining $O(\sqrt{T})$ rates for ECE is impossible. Recently, in a COLT 2024 paper, Qiao and Zheng [2024] showed that it was possible to make sequential predictions against an adversary guaranteeing expected distance to calibration growing at a rate of $O(\sqrt{T})$. Their algorithm is the solution to a minimax problem of size doubly-exponential in T. They leave as an open problem finding an explicit, efficient, deterministic algorithm for this problem. In this paper we resolve this problem, by giving an extremely simple such algorithm with an elementary analysis.

ALGORITHM 1.1. Almost-One-Step-Ahead

Input: Sequence of outcomes $y^{1:T} \in \{0,1\}^T$.
Output: Sequence of predictions $p^{1:T} \in \left\{0, \frac{1}{m}, \dots, 1\right\}^T$ for some discretization parameter $m > 0$.
for each $t = 1$ to T:

Given look-ahead predictions $\tilde{p}^{1:t-1}$, define the look-ahead bias conditional on a prediction p as:

$$\alpha_{\tilde{p}^{1:t-1}}(p) := \sum_{s=1}^{t-1} \mathbb{1}[\tilde{p}^s = p](\tilde{p}^s - y^s)$$

Choose two adjacent points $p_i = \frac{i}{m}, p_{i+1} = \frac{i+1}{m}$ satisfying:

$$\alpha_{\tilde{p}^{1:t-1}}(p_i) \le 0 \text{ and } \alpha_{\tilde{p}^{1:t-1}}(p_{i+1}) \ge 0$$

Arbitrarily predict $p^t = p_i$ or $p^t = p_{i+1}$.

Upon observing the outcome y^t, set the look-ahead prediction:

$$\tilde{p}^t = \text{argmin}_{p \in \{p_i, p_{i+1}\}} |p - y^t|$$

---

2 Setting

We study a sequential binary prediction setting: at every round t, a forecaster makes a prediction $p^t \in [0,1]$, after which an adversary reveals an outcome $y^t \in \{0,1\}$. Given a sequence of predictions $p^{1:T}$ and outcomes $y^{1:T}$, we measure expected calibration error (ECE) as follows:

$$\text{ECE}(p^{1:T}, y^{1:T}) = \sum_{p \in [0,1]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t) \right|$$

Following Qiao and Zheng [2024], we define *distance to calibration* to be the minimum ℓ_1 distance between a sequence of predictions produced by a forecaster and any *perfectly calibrated* sequence of predictions:

$$\text{CalDist}(p^{1:T}, y^{1:T}) = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T}) = \{q^{1:T} : \text{ECE}(q^{1:T}, y^{1:T}) = 0\}$ is the set of predictions that are perfectly calibrated against outcomes $y^{1:T}$. First we observe that distance to calibration is upper bounded by ECE.

LEMMA 2.1. (QIAO AND ZHENG [2024]) *Fix a sequence of predictions* $p^{1:T}$ *and outcomes* $y^{1:T}$. *Then,* $\mathrm{CalDist}(p^{1:T}, y^{1:T}) \le \mathrm{ECE}(p^{1:T}, y^{1:T})$.

Proof. For any prediction $p \in [0,1]$, define

$$\overline{y}^T(p) = \sum_{t=1}^{T} \frac{\mathbb{1}[p^t = p]}{\sum_{t=1}^{T} \mathbb{1}[p^t = p]} y^t$$

to be the average outcome conditioned on the prediction p. Consider the sequence $q^{1:T}$ where $q^t = \overline{y}^T(p^t)$. Observe that $q^{1:T}$ is perfectly calibrated. Thus, we have that

$$\begin{aligned}
\mathrm{CalDist}(p^{1:T}, y^{1:T}) &\le \|p^{1:T} - q^{1:T}\|_1 \\
&= \sum_{t=1}^{T} |p^t - q^t| \\
&= \sum_{p\in[0,1]} \sum_{t=1}^{T} \mathbb{1}[p^t = p] |p - \overline{y}^T(p)| \\
&= \sum_{p\in[0,1]} |p - \overline{y}^T(p)| \sum_{t=1}^{T} \mathbb{1}[p^t = p] \\
&= \sum_{p\in[0,1]} \left| p \sum_{t=1}^{T} \mathbb{1}[p^t = p] - \overline{y}^T(p) \sum_{t=1}^{T} \mathbb{1}[p^t = p] \right| \\
&= \sum_{p\in[0,1]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p - y^t) \right| \\
&= \mathrm{ECE}(p^{1:T}, y^{1:T})
\end{aligned}$$

□

The upper bound is not tight, however. The best known sequential prediction algorithm obtains ECE bounded by $O(T^{2/3})$ [Foster and Vohra, 1998], and it is known that there is no algorithm guaranteeing ECE below $O(T^{0.54389})$ [Qiao and Valiant, 2021, Dagan et al., 2024]. Qiao and Zheng [2024] give an algorithm that is the solution to a game of size doubly-exponential in T that obtains expected distance to calibration $O(\sqrt{T})$. Here we give an elementary analysis of a simple efficient deterministic algorithm (Algorithm 1.1) that obtains distance to calibration $2\sqrt{T} + 1$.

THEOREM 2.1. *Algorithm 1.1 (Almost-One-Step-Ahead) guarantees that against any sequence of outcomes,* $\mathrm{CalDist}(p^{1:T}, y^{1:T}) \le 2\sqrt{T} + 1$.

3 Analysis of Algorithm 1.1

Before describing the algorithm, we introduce some notation. We will make predictions that belong to a grid. Let $B_m = \{0, 1/m, ..., 1\}$ denote a discretization of the prediction space with discretization parameter $m > 0$, and let $p_i = i/m$. For a sequence of predictions $\tilde{p}^1, ..., \tilde{p}^t$ and outcomes $y^1, ..., y^t$, we define the bias conditional on a prediction p as:

$$\alpha_{\tilde{p}^{1:t}}(p) = \sum_{s=1}^{t} \mathbb{1}[\tilde{p}^s = p](\tilde{p}^s - y^s)$$

To understand our algorithm, it will be helpful to first state and analyze a hypothetical "lookahead" algorithm that we call "One-Step-Ahead", which is closely related to the algorithm and analysis given by Gupta and Ramdas [2022] in a different model. One-Step-Ahead produces predictions $\tilde{p}^1, ..., \tilde{p}^T$ as follows. At round t, before observing y^t, the algorithm fixes two predictions p_i, p_{i+1} satisfying $\alpha_{\tilde{p}^{1:t-1}}(p_i) \le 0$ and $\alpha_{\tilde{p}^{1:t-1}}(p_{i+1}) \ge 0$. Such a pair is guaranteed to exist, because by construction, it must be that for any history, $\alpha_{\tilde{p}^{1:t-1}}(0) \leq 0$ and $\alpha_{\tilde{p}^{1:t-1}}(1) \geq 0$. Note that a well known randomized algorithm obtaining diminishing ECE (and smooth calibration error) uses the same observation to carefully *randomize* between two such adjacent predictions [Foster, 1999, Foster and Hart, 2018]. Upon observing the outcome y^t, the algorithm outputs prediction $\tilde{p}^t = \text{argmin}_{p \in \{p_i, p_{i+1}\}} |p - y^t|$. Naturally, we cannot implement this algorithm, as it chooses its prediction only after observing the outcome, but our analysis will rely on a key property this algorithm maintains—namely, that it always produces a sequence of predictions with ECE upper bounded by $m + 1$, the number of elements in the discretized prediction space.

THEOREM 3.1. *For any sequence of outcomes, One-Step-Ahead achieves* $\text{ECE}(\tilde{p}^{1:T}, y^{1:T}) \leq m + 1$.

Proof. We will show that for any $p_i \in B_m$, we have $|\alpha_{\tilde{p}^{1:T}}(p_i)| \leq 1$, after which the bound on ECE will follow: $\text{ECE}(\tilde{p}^{1:T}, y^{1:T}) = \sum_{p_i \in B_m} |\alpha_{\tilde{p}^{1:T}}(p_i)| \leq m+1$. We proceed via an inductive argument. Fix a prediction $p_i \in B_m$. At the first round t_1 in which p_i is output by the algorithm, we have that $|\alpha_{\tilde{p}^{1:t_1}}(p_i)| = |p^{t_1} - y^{t_1}| \leq 1$. Now suppose after round $t-1$, we satisfy $|\alpha_{\tilde{p}^{1:t-1}}(p_i)| \leq 1$. If p_i is the prediction made at round t, it must be that either: $\alpha_{\tilde{p}^{1:t-1}}(p_i) \leq 0$ and $p_i - y^t \geq 0$; or $\alpha_{\tilde{p}^{1:t-1}}(p_i) \geq 0$ and $p_i - y^t \leq 0$. Thus, since $\alpha_{\tilde{p}^{1:t-1}}(p_i)$ and $p_i - y^t$ either take value 0 or differ in sign, we can conclude that

$$|\alpha_{\tilde{p}^{1:t}}(p_i)| = |\alpha_{\tilde{p}^{1:t-1}}(p_i) + p_i - y^t| \leq \max\{|\alpha_{\tilde{p}^{1:t-1}}(p_i)|, |p_i - y^t|\} \leq 1$$

which proves the theorem. □

Algorithm 1.1 (Almost-One-Step-Ahead) maintains the same state $\alpha_{\tilde{p}^{1:t}}(p)$ as One-Step-Ahead (which it can compute at round t after observing the outcome y_{t-1}). In particular, it does not keep track of the bias of its own predictions, but rather keeps track of the bias of the predictions that One-Step-Ahead *would have made*. Thus it can determine the pair p_i, p_{i+1} that One-Step-Ahead would commit to predict at round t. It cannot make the same prediction as One-Step-Ahead (as it must fix its prediction before the label is observed) — so instead it deterministically predicts $p^t = p_i$ (or $p^t = p_{i+1}$ — the choice can be arbitrary and does not affect the analysis). Since we have that $|p_i - p_{i+1}| \leq \frac{1}{m}$, it must be that for whichever choice One-Step-Ahead would have made, we have $|\tilde{p}^t - p^t| \leq \frac{1}{m}$. In other words, although Almost-One-Step-Ahead does not make the same predictions as One-Step-Ahead, it makes predictions that are within ℓ_1 distance T/m after T rounds. The analysis then follows by the ECE bound of One-Step-Ahead, the triangle inequality, and choosing $m = \sqrt{T}$.

Proof of Theorem 2.1. Observe that internally, Algorithm 1.1 maintains the sequence $\tilde{p}^1, ..., \tilde{p}^t$ which corresponds exactly to predictions made by One-Step-Ahead. Thus, by Lemma 2.1 and Theorem 3.1, we have that $\text{CalDist}(\tilde{p}^{1:T}, y^{1:T}) \leq \text{ECE}(\tilde{p}^{1:T}, y^{1:T}) \leq m+1$. Then, we can compute the distance to calibration of the sequence $p^1, ..., p^T$:

$$\begin{aligned}
\text{CalDist}(p^{1:T}, y^{1:T}) &= \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1 \\
&= \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - \tilde{p}^{1:T} + \tilde{p}^{1:T} - q^{1:T}\|_1 \\
&\leq \|p^{1:T} - \tilde{p}^{1:T}\|_1 + \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|\tilde{p}^{1:T} - q^{1:T}\|_1 \\
&\leq \frac{T}{m} + m + 1
\end{aligned}$$

where in the last step we use the fact that $|p^t - \tilde{p}^t| \leq 1/m$ for all t and thus $\|p^{1:T} - \tilde{p}^{1:T}\|_1 \leq T/m$. The result then follows by setting $m = \sqrt{T}$. □

Acknowledgements This work was supported in part by the Simons Collaboration on the Theory of Algorithmic Fairness, NSF grants FAI-2147212, CCF-2217062, CCF-1910534 and CCF- 2045128, an AWS AI Gift for Research on Trustworthy AI, and the Hans Sigrist Prize.

References

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1727–1740, 2023.

Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, Noah Golowich, Robert Kleinberg, and Princewill Okoroafor. Improved bounds for calibration via stronger sign preservation games, 2024. URL `https://arxiv.org/abs/2406.13668`.

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

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

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998. ISSN 00063444. URL `http://www.jstor.org/stable/2337364`.

Chirag Gupta and Aaditya Ramdas. Faster online calibration without randomization: interval forecasts and the power of two choices. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4283–4309. PMLR, 02–05 Jul 2022. URL `https://proceedings.mlr.press/v178/gupta22b.html`.

Mingda Qiao and Gregory Valiant. Stronger calibration lower bounds via sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 456–466, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3451050. URL `https://doi.org/10.1145/3406325.3451050`.

Mingda Qiao and Letian Zheng. On the distance from calibration in sequential prediction. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4307–4357. PMLR, 30 Jun–03 Jul 2024. URL `https://proceedings.mlr.press/v247/qiao24a.html`.