# A Perfectly Truthful Calibration Measure

**Jason Hartline** HARTLINE@NORTHWESTERN.EDU
*Computer Science Department, Northwestern University*
**Lunjia Hu** LUNJIA@ALUMNI.STANFORD.EDU
*Khoury College of Computer Sciences, Northeastern University*
**Yifan Wu** YIFAN.WU2357@GMAIL.COM
*Microsoft Research, New England*

**Editors:** Steve Hanneke and Tor Lattimore

## Abstract

Calibration requires that predictions are conditionally unbiased and, therefore, reliably interpretable as probabilities. A calibration measure quantifies how far a predictor is from perfect calibration. As introduced by Haghtalab et al. (2024), a calibration measure is truthful if it is minimized in expectation when a predictor outputs the ground-truth probabilities. Predicting the true probabilities guarantees perfect calibration, but in reality, when calibration is evaluated on a random sample, all known calibration measures incentivize predictors to lie in order to appear more calibrated. This lack of truthfulness motivated Haghtalab et al. (2024) and Qiao and Zhao (2025) to construct *approximately* truthful calibration measures in the sequential prediction setting, but no *perfectly* truthful calibration measure was known to exist even in the more basic batch setting.

We design a simple, perfectly and strictly truthful, sound, and complete calibration measure in the batch setting: Averaged Two-Bin Calibration Error (ATB). ATB is quadratically related to two existing calibration measures: the smooth calibration error (smCAL) and the lower distance to calibration (distCAL). The simplicity of our definition of ATB makes it efficient and straightforward to compute, allowing us to give the first linear-time calibration testing algorithm, improving a result of Hu et al. (2024). We also introduce a general recipe for constructing truthful measures based on the variance additivity of independent random variables, which proves the truthfulness of ATB as a special case and allows us to construct other truthful calibration measures, such as quantile-binned $\ell_2$ Expected Calibration Error (ECE).

**Keywords:** Calibration error, truthfulness.

## 1. Introduction

Probabilistic forecasting has become increasingly important in modern AI-assisted decision-making. Unlike deterministic classification, probabilistic forecasts provide uncertainty quantification, allowing assessment of risks. One desired property of probabilistic predictions is *calibration*, which requires predictions to be conditionally unbiased and, therefore, reliably interpretable as probabilities. For example, neural networks for tumor diagnosis are trained to output a prediction $r \in [0, 1]$, ideally interpretable as the expectation of a binary state $y \in \{0, 1\}$: the tumor segment being malignant or not. The neural network is calibrated if, conditioned on the output $r$ being, say $40\%$, the probability that the tumor is malignant $\Pr[y = 1|r = 40\%]$ is also $40\%$.

A calibration measure quantifies how far a predictor is from perfect calibration. The Expected Calibration Error (ECE) is a canonical calibration measure proposed by Foster and Vohra (1997). Given an empirical distribution of predictions and states, if conditioned on a reported prediction $r \in [0, 1]$, the actual empirical frequency of the state $y = 1$ is $\widehat{r} := \Pr[y = 1|r]$, then the absolute

bias in prediction is $|r - \widehat{r}|$. The Expected Calibration Error (ECE) is defined as the expected bias in predictions, $\mathbb{E}_r[|r - \widehat{r}|]$.

We study the truthfulness of calibration measures, following prior work (Haghtalab et al., 2024; Qiao and Zhao, 2025). An error measure is truthful if it incentivizes a predictor to output the truth, i.e., the expected error is minimized when the predictor reports the true probabilities. However, no known calibration measure is truthful. Even a miscalibrated predictor can have lower expected error than the truthful predictor when evaluated by known calibration measures. We explain this non-truthfulness in Section 1.1. The main result of this paper is a perfectly truthful calibration error for the batch setting.

### 1.1. Non-truthfulness of Known Calibration Measures

Both very informative predictors and very uninformative predictors can be perfectly calibrated. A well-known strategy to make an informative but miscalibrated predictor seem more calibrated is to average its predictions (DeGroot and Fienberg, 1983). This distorts the informativeness of the predictor but can cause calibration errors to cancel and, thus, improve. All prior notions of calibration error are manipulable by the simple and extreme version of this strategy: predict the base rate.

Example 1 illustrates the non-truthfulness of ECE. We consider the batch setting: a sample of $T$ individuals whose binary states $\boldsymbol{y} = (y_1, \ldots, y_T) \in \{0,1\}^T$ are independently drawn from the Bernoulli distributions with means $\boldsymbol{p} = (p_1, \ldots, p_T) \in [0,1]^T$ (denoted by $\boldsymbol{y} \sim \boldsymbol{p}$). We say $\boldsymbol{p} = (p_1, \ldots, p_T)$ are the ground-truth probabilities. In the example below, a predictor strictly benefits from reporting the base rate of the ground-truth probabilities.

**Example 1 (ECE is not truthful, c.f. Qiao and Valiant, 2021)** *Suppose the ground-truth probabilities are $\boldsymbol{p} = (p_1, \ldots, p_T)$ where each $p_t$ is distributed independently and uniformly from $[1/3, 2/3]$. An uninformative predictor that always predicts $r_1 = \cdots = r_T = 0.5$ achieves an expected empirical* ECE $= O\left(\sqrt{1/T}\right)$*, the sampling error. However, a truthful predictor who reports $r_t = p_t$ results in a higher empirical* ECE $\geq 1/3$*. This is because the predictions $r_1, \ldots, r_T \in [1/3, 2/3]$ are almost surely distinct, so the empirical conditional expectation $\widehat{r}_t := \mathbb{E}_{(r,y)\sim \mathrm{Unif}((r_t, y_t)_{t\in[T]})}[y | r = r_t]$ is simply $y_t \in \{0,1\}$, giving $|r_t - \widehat{r}_t| = |r_t - y_t| \geq 1/3$.*

Example 1 shows that the obvious uninformative prediction achieves a lower ECE. Even worse, there are *miscalibrated* predictors (e.g. the predictor that always predicts $0.5 + \epsilon$ for a small $\epsilon > 0$) achieving smaller ECE than the *calibrated* truthful predictor. Thus ECE do not rank predictors correctly based on how calibrated they are.

Predicting the uninformative base rate incurs a lower calibration error for all known calibration measures. It happens not just for ECE in Example 1 and its variants[1], but also for continuous calibration measures[2] such as the smooth calibration error (Kakade and Foster, 2008), the distance to calibration (Błasiok et al., 2023) and its variants, etc, irrespective of the sample size $T$. Moreover, it happens consistently across *every* realization of the states $\boldsymbol{y}$, *not just in expectation*. Specifically,

1. Variants of ECE include $\ell_\alpha$-ECE, where we replace the absolute bias $|r - \widehat{r}|$ with $|r - \widehat{r}|^\alpha$ for an arbitrary $\alpha \geq 1$, as well as binned versions of $\ell_\alpha$-ECE.
2. These are calibration measures that are continuous as a function of the predictions. Note that ECE and binned ECE are not continuous.

for *every* realization of $\boldsymbol{y} \in \{0,1\}^T$ and *every* prediction sequence $\boldsymbol{r} = (r_1, \ldots, r_T) \in [0,1]^T$, all these calibration measures CAL satisfy

$$\text{CAL}(\boldsymbol{r}, \boldsymbol{y}) \geq \text{CAL}(\bar{\boldsymbol{r}}, \boldsymbol{y}), \tag{1}$$

where $\bar{\boldsymbol{r}} = (\bar{r}, \ldots, \bar{r})$ is the constant predictor that always predicts the average $\bar{r} := \frac{1}{T}\sum_{t=1}^{T} r_t$. This obvious and uninformative strategy always achieves (weakly) lower calibration error. For many realizations of states $\boldsymbol{y}$, the error is strictly lower. We formally prove this observation in Theorem 45.

### 1.2. Our Goal: Perfectly Truthful Calibration Measures

Measuring and optimizing for calibration **non-truthfully** only makes the predictions **less trustworthy**, going in the very opposite direction of the goal of calibration. Recall the definition of truthfulness: a calibration measure CAL is truthful if for every sequence of ground-truth probabilities $\boldsymbol{p} \in [0,1]^T$, the expected empirical calibration error $\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\text{CAL}(\boldsymbol{r}, \boldsymbol{y})]$ of predictions $\boldsymbol{r} \in [0,1]^T$ on a random sample $\boldsymbol{y} \sim \boldsymbol{p}$ is minimized when $\boldsymbol{r} = \boldsymbol{p}$. From a machine learning perspective, a truthful measure helps identify the Bayes optimal predictor (Gneiting, 2011) because it correctly ranks ground-truth predictions with the lowest expected error. From a game-theoretic perspective, a truthful measure incentivizes an optimizing predictor to output their true beliefs, where we view $\boldsymbol{p}$ as the predictor's subjective belief about the probabilities, which might differ from the true probabilities. For the tumor risk prediction task, if assessed by a non-truthful calibration measure, a doctor is incentivized to report a prediction different from their true probabilistic assessment of the tumor risk, to make the predictions look more calibrated. Such an incentivized misreport can hardly be trusted.

We focus on perfect truthfulness in the batch setting, where all states are independently drawn. Previous work (Haghtalab et al., 2024; Qiao and Zhao, 2025) designs approximately truthful calibration measures in the sequential prediction setting, where the states $y_1, \ldots, y_T$ are revealed sequentially after each prediction $r_t$ is made. We observe that, in the simpler batch setting, some existing measures are approximately truthful, such as the smooth calibration error (Kakade and Foster, 2008; Błasiok et al., 2023) and the calibration measures proposed by Haghtalab et al. (2024) and Qiao and Zhao (2025). Yet, no known calibration measure is perfectly truthful.

The two minimum requirements of a calibration measure are completeness and soundness. At the population level, a calibration measure should be zero on every calibrated prediction-state distribution and positive on every miscalibrated one (Definitions 11 and 12). In applications, we also need an empirical estimator that converges to this population measure; for Averaged Two-Bin Calibration Error (ATB) and its $\ell_1$ variant, this estimator is the plug-in estimator that evaluates the same measure on the empirical distribution of the sample (Definition 13). Haghtalab et al. (2024) point out that some error metrics, such as the well-known squared error $\frac{1}{T}\sum_t (r_t - y_t)^2$, are truthful but far from being a complete and sound calibration measure. The squared error of a calibrated predictor may remain bounded away from zero.

Our main result shows that truthfulness can be achieved via surprisingly simple constructions in the batch setting, while preserving the completeness and soundness of existing calibration measures.

## 2. Truthfulness of Binned Errors

We now explain our construction of truthful calibration measures. The central object is a general recipe for constructing truthful binning-based errors, *Unnormalized Binned Squared Errors* (UBSEs). We first give the intuition behind Unnormalized Binned Squared Errors (UBSEs), then list quantile-binned $\ell_2$ Expected Calibration Error (ECE) as a simple example, and finally state Averaged Two-Bin Calibration Error (ATB) as the main Unnormalized Binned Squared Error (UBSE) used throughout the paper. The formal appendix treatment is in Appendices B and C.

### 2.1. Lemma 25: Truthfulness from Variance Additivity

We discuss the idea behind our construction of a general family of truthful measures, i.e., Unnormalized Binned Squared Errors (UBSEs). As mentioned earlier, Averaged Two-Bin Calibration Error (ATB) is a member of this family, so its truthfulness follows as a consequence.

Our goal is to measure the calibration error of a sequence of predictions $\boldsymbol{r} = (r_1, \ldots, r_T) \in [0,1]^T$ on the states $\boldsymbol{y} = (y_1, \ldots, y_T) \in \{0,1\}^T$. Here, each state $y_t$ is sampled independently from the Bernoulli distribution with mean $p_t \in [0,1]$, where $\boldsymbol{p} = (p_1, \ldots, p_T) \in [0,1]^T$ are the true probabilities (denoted by $\boldsymbol{y} \sim \boldsymbol{p}$).

**Standard binned $\ell_2$ Expected Calibration Error (ECE) is not truthful.** If we divide the predictions $r_t$ into bins based on the *indices* $t \in [T]$ rather than the *values* $r_t \in [0,1]$, then truthfulness can be easily achieved by the $\ell_2$ version of binned Expected Calibration Error (ECE), $\ell_2$-BINECE. Concretely, consider a fixed partition $\mathcal{B} = (B_1, \ldots, B_k)$ of the index space $[T]$ into bins: $[T] = B_1 \cup \cdots \cup B_k$. The $\ell_2$-BINECE follows the standard computation of Expected Calibration Error (ECE) but replacing the $\ell_1$ error with squared error:

$$\begin{aligned}\ell_2\text{-BINECE}_{\mathcal{B}}(\boldsymbol{r}, \boldsymbol{y}) &= \sum_{i \in [k]} \underbrace{\frac{|B_i|}{T}}_{\text{weigh by fraction}} \cdot \left( \underbrace{\frac{1}{|B_i|}}_{\text{normalize by size}} \underbrace{\sum_{t \in B_i} (r_t - y_t)}_{\text{the bias in } B_i} \right)^2 \\ &= \sum_{i \in [k]} \frac{1}{T|B_i|} \left( \sum_{t \in B_i} (r_t - y_t) \right)^2 . \end{aligned} \tag{2}$$

Assuming the index partition $\mathcal{B}$ is fixed, the truthfulness of $\ell_2$-BINECE comes from the truthfulness of squared error: within each bin $B_i$, the expected squared bias over $\boldsymbol{y} \sim \boldsymbol{p}$

$$\mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}} \left[ \left( \sum_{t \in B_i} (r_t - y_t) \right)^2 \right] \tag{3}$$

is minimized if and only if $\sum_{t \in B_i} r_t = \sum_{t \in B_i} p_t$, which is implied by predicting the truth $r_t = p_t$. However, $\ell_2$-BINECE in its standard form is not truthful because it bins by prediction values $r_t$ rather than by fixed indices. It first partitions the prediction space $[0,1]$ into intervals and then assigns each index $t$ to the interval containing $r_t$. A strategic predictor can therefore manipulate the partition $\mathcal{B}$ by making untruthful predictions. In the same spirit as Example 1, always predicting

the same value $r_1 = \cdots = r_T$ puts all indices in one bin, producing a different index partition $\mathcal{B}'$ than truthful prediction. This manipulated coarser partition can significantly reduce the expected $\ell_2$-BINECE because of the bin-size-based normalization $1/|B_i|$ in (2).

**Unnormalized Binned Squared Errors (UBSEs) are truthful.** The unnormalized version removes the bin-size-based normalization $1/|B_i|$ from (2). For a fixed partition $\mathcal{B} = (B_1, \ldots, B_k)$, define

$$\textsc{Cal}_{\mathcal{B}}(\boldsymbol{r}, \boldsymbol{y}) := \frac{1}{T^2} \sum_{i=1}^{k} \left( \sum_{t \in B_i} (r_t - y_t) \right)^2 .$$

The example above suggests that truthfulness can be restored if the expected error of truthful predictions is *invariant* to the index partition. For any partition $\mathcal{B}'$ that could be induced by a strategic report $\boldsymbol{r}$, the fixed-partition squared-error argument gives

$$\mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}}[\textsc{Cal}_{\mathcal{B}'}(\boldsymbol{p}, \boldsymbol{y})] \leq \mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}}[\textsc{Cal}_{\mathcal{B}'}(\boldsymbol{r}, \boldsymbol{y})]. \tag{4}$$

To establish truthfulness for a report-dependent binning rule, we also need to compare the truthful predictions on their own partition $\mathcal{B}$ against the strategic report on $\mathcal{B}'$:

$$\mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}}[\textsc{Cal}_{\mathcal{B}}(\boldsymbol{p}, \boldsymbol{y})] \leq \mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}}[\textsc{Cal}_{\mathcal{B}'}(\boldsymbol{r}, \boldsymbol{y})]. \tag{5}$$

Thus, it suffices for the truthful expected error to be invariant to the partition:

$$\mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}}[\textsc{Cal}_{\mathcal{B}'}(\boldsymbol{p}, \boldsymbol{y})] = \mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}}[\textsc{Cal}_{\mathcal{B}}(\boldsymbol{p}, \boldsymbol{y})] \quad \text{for any partitions } \mathcal{B}, \mathcal{B}'. \tag{6}$$

Unnormalized Binned Squared Errors (UBSEs) achieve this invariance. For truthful predictions ($r_t = p_t$), the expected squared bias in bin $B_i$ is

$$\mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}} \left[ \left( \sum_{t \in B_i} (p_t - y_t) \right)^2 \right] = \textsc{Var}_{\boldsymbol{y} \sim \boldsymbol{p}} \left[ \sum_{t \in B_i} y_t \right] = \sum_{t \in B_i} \textsc{Var}_{\boldsymbol{y} \sim \boldsymbol{p}} [y_t] = \sum_{t \in B_i} p_t(1 - p_t), \tag{7}$$

where the second equality uses variance additivity for the independent states $y_t$. Summing (7) over the bins without any bin-size-based normalization gives

$$\mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}} [\textsc{Cal}_{\mathcal{B}}(\boldsymbol{p}, \boldsymbol{y})] = \frac{1}{T^2} \sum_{t \in [T]} p_t(1 - p_t),$$

which is independent of the partition $\mathcal{B}$ and satisfies (6). Therefore, even if a strategic report induces a different partition, truth is still at least as good on that same partition and the truthful benchmark is the same across partitions. This is the variance-additivity argument behind Unnormalized Binned Squared Errors (UBSEs), formalized in Lemma 25 and theorem 24. Moreover, any calibrated finite prediction sequence achieves expected error $\frac{1}{T^2} \sum_{t \in [T]} p_t(1 - p_t) = O(1/T)$ (see Lemma 25), giving a vanishing empirical-error guarantee as $T \to \infty$. The global normalization by $1/T^2$ ensures this scaling without affecting truthfulness, unlike the per-bin normalization in (2).

**Note on randomized binning.** The same analysis allows the partition $\mathcal{B}$ to be randomized: condition on the realized partition, apply the fixed-partition argument above, and then average over the partition randomness. This gives the full family of Unnormalized Binned Squared Errors (UBSEs) in Appendix B, where the possibly randomized binning strategy may depend on the reports $r_1, \ldots, r_T$ but not on the realized states. This family is not automatically sound; soundness depends on the binning strategy. Below, we present a sound and complete calibration error.

**Example: Quantile-binned $\ell_2$-BINECE is truthful.** As an example of an Unnormalized Binned Squared Error (UBSE), simple modifications make the $\ell_2$ binned Expected Calibration Error (ECE), $\ell_2$-BINECE, truthful by binning predictions according to quantiles. With $k$ bins, the following Unnormalized Binned Squared Error (UBSE) is truthful.

- Sort the samples by reported predictions with $r_1 \leq \cdots \leq r_T$. Break ties uniformly at random.
- Divide predictions into $k$ bins, with $\frac{T}{k}$ predictions in each bin.
- Calculate UBSE.

Binning according to quantiles ensures that each bin contains the same number of predictions and thus, the normalization factors based on bin sizes $|B_i|$ in (2) no longer break truthfulness.

### 2.2. Averaged Two-Bin Calibration Error (ATB) as a Special Case

The proposed calibration measure in this paper is Averaged Two-Bin Calibration Error (ATB), which uses the simplest nontrivial randomized binning scheme: choose one threshold $q$ uniformly from $[0, 1]$ and use the two bins $[0, q)$ and $[q, 1]$.

**Definition 1 (Averaged two-bin calibration error)** *Given predictions $\boldsymbol{r} = (r_1, \ldots, r_T) \in [0, 1]^T$ and states $\boldsymbol{y} = (y_1, \ldots, y_T) \in \{0, 1\}^T$, we define*

$$\mathrm{ATB}(\boldsymbol{r}, \boldsymbol{y}) := \mathbb{E}_{q \sim \mathrm{Unif}([0,1])} \left[ \frac{1}{T^2} \left( \left( \sum_{t: r_t < q} (r_t - y_t) \right)^2 + \left( \sum_{t: r_t \geq q} (r_t - y_t) \right)^2 \right) \right].$$

For every fixed threshold $q$, the two threshold bins form an Unnormalized Binned Squared Error (UBSE) partition that depends only on the reported predictions. Averaging over $q$ preserves the UBSE structure, so Averaged Two-Bin Calibration Error (ATB) is truthful by the UBSE error decomposition (Lemma 25 and theorem 24). The formal statement in Theorem 35 also records the inherited decomposition for ATB, and Theorem 26 gives strict ex-ante truthfulness.

**Consequences of truthfulness.** Truthfulness also gives evaluation guarantees that are not tied to the specific two-bin construction. A follow-up work to our paper shows that a truthful calibration measure preserves the Blackwell ordering among calibrated predictors and weakly improves when a miscalibrated predictor is recalibrated (Lu et al., 2025). Thus Averaged Two-Bin Calibration Error (ATB) and quantile-binned $\ell_2$ Expected Calibration Error (ECE) inherit the qualitative robustness expected from proper, truthful evaluation rules, while still targeting calibration rather than prediction loss directly.

## 3. Properties and Consequences of ATB

Truthfulness is only the first requirement on Averaged Two-Bin Calibration Error (ATB). In this section, we present the remaining guarantees: ATB is complete and sound as a distributional calibration measure, its plug-in estimator is consistent, it is continuous in the predictions, it has optimal sample complexity, and it can be computed or approximated efficiently. The formal statements and proofs are in the appendices, and we keep the references to those later results explicit.

### 3.1. Completeness, Soundness, and Estimator Consistency

We show that ATB is complete and sound given distributional access to the joint distribution over prediction and states. Moreover, the sample estimator is consistent and converges to the distributional ATB when sample size grows. The completeness, soundness, and estimator consistency guarantees are summarized in Theorem 30. Their proof uses the $\ell_1$ two-bin proxy $\ell_1$-ATB and the relationship in Corollary 32: $\ell_1$-ATB is a constant-factor approximation to both smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL), while ATB is quadratically related to $\ell_1$-ATB by Lemma 29. This yields the quadratic relationship between Averaged Two-Bin Calibration Error (ATB) and the standard complete and sound calibration measures smooth calibration error (SMCAL) (Kakade and Foster, 2008) and lower distance to calibration (DISTCAL) (Błasiok et al., 2023).

**Definition 2** *For prediction and state sequences $\boldsymbol{r}, \boldsymbol{y}$, the $\ell_1$ variant of Averaged Two-Bin Calibration Error (ATB) is*

$$\ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}) := \mathbb{E}_{q\sim\text{Unif}([0,1])}\left[\frac{1}{T}\left(\left|\sum_{t:r_t<q}(r_t - y_t)\right| + \left|\sum_{t:r_t\geq q}(r_t - y_t)\right|\right)\right].$$

**Theorem 3 (Informal, Corollary 32)** *The proxy $\ell_1$-ATB is a constant-factor approximation to smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL):*

$$\frac{1}{3}\,\text{DISTCAL}(\boldsymbol{r}, \boldsymbol{y}) \leq \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}) \leq 3\,\text{DISTCAL}(\boldsymbol{r}, \boldsymbol{y}),$$
$$\frac{2}{3}\,\text{SMCAL}(\boldsymbol{r}, \boldsymbol{y}) \leq \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}) \leq 6\,\text{SMCAL}(\boldsymbol{r}, \boldsymbol{y}).$$

*Consequently,*

$$\frac{1}{18}\text{DISTCAL}(\boldsymbol{r}, \boldsymbol{y})^2 \leq \frac{2}{9}\text{SMCAL}(\boldsymbol{r}, \boldsymbol{y})^2 \leq \text{ATB}(\boldsymbol{r}, \boldsymbol{y}) \leq 3\text{DISTCAL}(\boldsymbol{r}, \boldsymbol{y}) \leq 6\text{SMCAL}(\boldsymbol{r}, \boldsymbol{y}). \tag{8}$$

Averaged Two-Bin Calibration Error (ATB) also has the regularity and efficiency properties needed for empirical use. Theorem 36 proves Lipschitz continuity under perturbations of the predictions. Theorem 37 shows that ATB and $\ell_1$-ATB can both be estimated to additive error $\varepsilon$ from $O(1/\varepsilon^2)$ independent and identically distributed (i.i.d.) samples. Finally, Theorem 39 gives an $O(T \log T)$ exact algorithm and an $O(T + 1/\varepsilon)$ additive-approximation algorithm.

### 3.2. Validity and Linear-Time Calibration Testing

The same structure also gives an efficient calibration tester. In Theorem 40, Averaged Two-Bin Calibration Error (ATB) is shown to be $O(1/\sqrt{T})$-valid for lower distance to calibration (DISTCAL) (and therefore for smooth calibration error (SMCAL) up to constants), and this rate is information-theoretically optimal. As a consequence, the calibration testing problem of Hu et al. (2024) can be solved by computing Averaged Two-Bin Calibration Error (ATB) on $T = O(1/\varepsilon^2)$ samples. The exact computation already improves the runtime to $O(T \log T)$, and the additive approximation from Theorem 39 gives a linear-time tester.

### 3.3. Technical Overview: Two-Bin Approximation of the Smooth Calibration Error (Corollary 32)

Our Unnormalized Binned Squared Error (UBSE) framework is flexible with regard to how the bins should be chosen (including how many bins should be chosen). However, it is not obvious to find an appropriate binning scheme and show that the corresponding UBSE is polynomially related to existing calibration error metrics such as smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL).

Our construction of Averaged Two-Bin Calibration Error (ATB) is quadratically related to smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL). As mentioned earlier, we prove this result by showing that $\ell_1$-ATB (Definition 2) gives a constant-factor approximation for SMCAL and DISTCAL. Here we explain the intuition behind this analysis.

Our analysis is divided into the following two results, showing the upper and lower bounds on $\ell_1$-ATB separately:

$$\text{Lemma 33:} \quad \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}) \leq 3\,\text{DISTCAL}(\boldsymbol{r}, \boldsymbol{y}) \tag{9}$$

$$\text{Lemma 34:} \quad \text{SMCAL}(\boldsymbol{r}, \boldsymbol{y}) \leq \frac{3}{2}\,\ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}). \tag{10}$$

The desired constant-factor approximation (Theorem 3) then follows from the previous result that SMCAL and DISTCAL are themselves constant-factor approximations of each other (Proposition 9) (Błasiok et al., 2023).

While neither inequality is straightforward to prove, the relatively more technically involved and, perhaps, more surprising direction is the latter inequality (10) showing that SMCAL can be upper-bounded by $\ell_1$-ATB up to a constant factor. Indeed, the intuition behind the previous notion of *interval calibration error* INTCAL (Błasiok et al., 2023) is that having too few bins tends to underestimate SMCAL, and if the calibration error is much smaller than the average bin width, we should increase the number of bins to faithfully capture SMCAL.[3] The reasoning is that having fewer bins makes more predictions fall into the same bin, among which the positive and negative biases $r_t - y_t$ cancel out, thus more likely to cause underestimation. For example, having only one bin gives the following UBSE:

$$\text{CAL}(\boldsymbol{r}, \boldsymbol{y}) = \left(\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)\right)^2,$$

3. Consequently, the number of bins used to define $\text{INTCAL}(\boldsymbol{r}, \boldsymbol{y})$ depends on both $\boldsymbol{r}$ and $\boldsymbol{y}$. In UBSE, the binning scheme can only depend on $\boldsymbol{r}$ in order for our truthfulness analysis to hold.

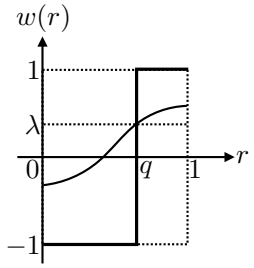


Figure 1: Writing $w$ as a convex combination of threshold functions.

which clearly underestimates smooth calibration error (SMCAL) (it can be zero even when $\boldsymbol{r}$ is mis-calibrated, in which case $\text{SMCAL}(\boldsymbol{r}, \boldsymbol{y})$ is always positive). Therefore, based on this previous intuition, it is somewhat surprising that having just two bins suffices to establish (10).

Here $\ell_1$-ATB denotes the $\ell_1$ variant of Averaged Two-Bin Calibration Error (ATB). Proving (10) is equivalent to showing that for any 1-Lipschitz weight function $w : [0, 1] \to [-1, 1]$,

$$\frac{1}{T} \sum_{t \in [T]} w(r_t) \cdot (r_t - y_t) \leq \frac{3}{2} \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}). \tag{11}$$

This equivalence follows from the definition of smooth calibration error (SMCAL): it is the supremum of the left-hand side over all 1-Lipschitz $w : [0, 1] \to [-1, 1]$ (Definition 8).

To illustrate our proof idea, let us first assume that the weight function $w$ is not only Lipschitz, but also monotonically increasing and differentiable (represented by the curve in Figure 1). The key observation is that we can write $w$ as a convex combination of threshold functions as follows. Take a random threshold $\lambda$ uniformly distributed from $[-1, 1]$ and consider the threshold function $w_\lambda(r) := \mathsf{sign}(w(r) - \lambda)$ (represented by the bold step function in Figure 1). That is, $w_\lambda(r) = 1$ if $w(r) \geq \lambda$, and $w_\lambda(r) = -1$ if $w(r) < \lambda$. The following key identity expresses $w$ as a convex combination of the threshold functions $w_\lambda$:

$$w(r) = \mathbb{E}_{\lambda \sim \text{Unif}([-1,1])}[w_\lambda(r)] \quad \text{for every } r \in [0, 1]. \tag{12}$$

Now for a fixed threshold $\lambda \in [-1, 1]$, let $q := w^{(-1)}(\lambda) \in [0, 1]$ be the corresponding threshold on the $r$-axis, where $w^{(-1)}$ is the inverse of $w$ (see Figure 1). In the boundary cases when $\lambda > w(1)$, we choose $q = 1$, and similarly, when $\lambda < w(0)$ we choose $q = 0$. This ensures[4]

$$w_\lambda(r) = \mathsf{sign}(r - q) \quad \text{for every } r \in [0, 1]. \tag{13}$$

Let $Q$ be the distribution of the resulting $q$ from $\lambda \sim \text{Unif}([-1, 1])$. By (12) and (13), we can rewrite the left-hand side of (11) as

$$\frac{1}{T} \sum_{t \in [T]} w(r_t) \cdot (r_t - y_t) = \mathbb{E}_{q \sim Q} \left[ \frac{1}{T} \sum_{t \in [T]} \mathsf{sign}(r_t - q) \cdot (r_t - y_t) \right]. \tag{14}$$

4. One tiny caveat which we ignore here is that when $\lambda > w(1)$ and thus $q = 1$, this identity does not hold at one point: $r = 1$.

For each fixed choice of $q$, it is straightforward to show that the quantity inside the expectation in (14) is upper-bounded by the $\ell_1$ variant of Averaged Two-Bin Calibration Error ($\ell_1$-ATB) at the same fixed bin threshold $q$ (Definition 2). However, the random variable $q$ is distributed differently in the two cases. It is drawn from the distribution $Q$ in (14), whereas it is uniformly distributed over $[0, 1]$ in the definition of $\ell_1$-ATB.

What remains is to relate the two distributions: $Q$ and $\mathrm{Unif}([0, 1])$. Recall that $q \sim Q$ is obtained as $q = w^{(-1)}(\lambda)$ for uniformly distributed $\lambda \in [-1, 1]$. It follows that the probability density function (PDF) of $q \sim Q$ is exactly the probability density function (PDF) of $\lambda$ (which is $1/2$ everywhere in $[-1, 1]$) times the derivative $\nabla w(q)$, except at the boundaries $q = 0, 1$. Since $w$ is 1-Lipschitz, we have $\nabla w(q) \leq 1$, and thus the PDF of $q \sim Q$ is at most $1/2$ everywhere in the open interval $(0, 1)$. This is sufficient to bound the expectation over $q \sim Q$ in (14) by the expectation over $\mathrm{Unif}([0, 1])$ in the definition of $\ell_1$-ATB (Definition 28). The boundary cases of $q = 0, 1$ need to be handled separately, but that turns out to be relatively straightforward.

To fully prove (10), we need to remove the monotonicity and differentiability assumptions on $w$, which is achieved by our formal proof in Appendix C.1.1. Roughly speaking, without monotonicity, the convex combination of the threshold functions that expresses $w$ might have negative coefficients (so it is a linear combination rather than a convex combination), but the absolute values of the coefficients can still be controlled using the Lipschitzness of $w$. The differentiability assumption can be removed by focusing on the finite set $\{r_1, \ldots, r_T\}$ rather than the full domain $[0, 1]$ of $w$.

## 4. Related Work

**Truthful Calibration Measures.** Previous work (Haghtalab et al., 2024; Qiao and Zhao, 2025) on approximate truthful calibration errors is closest to our paper. They design multiplicatively truthful calibration errors in the sequential prediction problem. In the sequential prediction setting, a sequence of $T$ potentially correlated states is drawn from a distribution. At each period, the predictor predicts, and one state is revealed. Our work studies a different batch setting where all $T$ states are independently drawn and revealed simultaneously after all predictions. An error metric is approximately truthful if predicting the true conditional probability of the next state is a constant approximation of the optimal strategy. Haghtalab et al. (2024) shows that subsampled smooth calibration error is multiplicatively truthful for the sequential prediction setting, implying the smooth calibration error is multiplicatively truthful for the batch setting. Qiao and Zhao (2025) shows that, in the sequential setting, there does not exist a perfectly truthful calibration error that upper-bounds the worst-case external regret for decision-makers. The impossibility in the sequential setting does not apply to our problem. It also remains open whether there exists a perfectly truthful calibration error metric for the sequential setting while satisfying other completeness and soundness properties.

**Calibration Measures.** Foster and Vohra (1997) first proposed the Expected Calibration Error (ECE). The binned Expected Calibration Error (ECE) serves as a widely-used empirical proxy of ECE (Guo et al., 2017; Minderer et al., 2021). Kleinberg et al. (2023) observes that, if predictions are used for downstream decision-making, Expected Calibration Error (ECE) upper-bounds the swap regret of any downstream decision-maker. Following the decision-making purpose of calibration, Hu and Wu (2024) proposes Calibration Decision Loss (CDL), the worst-case swap regret of any normalized downstream decision task, and shows Calibration Decision Loss (CDL) is quantitatively different from ECE. Okoroafor et al. (2025) introduce the notion of proper calibration as a key ingredient for designing improved algorithms for omniprediction (Gopalan et al., 2022, 2023).

Błasiok et al. (2023) introduced the distance to calibration. In their framework, a calibration error is consistent if it is polynomially related to the distance to calibration. They showed that the smooth calibration error (Kakade and Foster, 2008) and the Laplace kernel calibration error (Kumar et al., 2018) are both consistent, and introduced a binning-based consistent calibration error: the interval calibration error.

**Proper Scoring Rules (a.k.a. truthful losses).** Initiated by McCarthy (1956); Savage (1971), extensive work focused on the characterization of proper scoring rules, the class of truthful loss functions. Lambert (2011) characterizes elicitable statistics of a distribution, for example, the mean of a distribution, not the variance of a distribution. Winkler et al. (1996) provides proper scoring rules for the confidence interval, and Frongillo and Kash (2014) provides a characterization of proper scoring rules for eliciting linear properties. Li et al. (2022) gives computational results of proper scoring rules.

## 5. Organization of Appendices

The appendices of the paper are organized as follows. Appendix A establishes the basic setup, including the definitions of existing calibration errors (Appendix A.1), completeness, soundness, and estimator consistency (Appendix A.2), the validity of calibration errors via calibration tests (Appendix A.4), and the truthfulness of calibration errors (Appendix A.3). In Appendix B, we introduce the Unnormalized Binned Squared Errors (UBSEs), a general family of truthful binning-based error metrics. In Appendix C, we introduce our proposed calibration error, the Averaged Two-Bin Calibration Error (ATB), as a special case of an Unnormalized Binned Squared Error (UBSE) and prove its truthfulness, continuity, sample efficiency, and computational efficiency. In Appendix C.1.1, we prove the quadratic relationship between ATB and the existing calibration errors smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL) by showing that $\ell_1$-ATB is a constant-factor approximation of SMCAL and DISTCAL.

## 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.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40–55, 1997.

Rafael Frongillo and Ian Kash. General truthfulness characterizations via convex analysis. In *Web and Internet Economics: 10th International Conference, WINE 2014, Beijing, China, December 14-17, 2014. Proceedings 10*, pages 354–370. Springer, 2014.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:21, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-217-4. doi: 10.4230/LIPIcs.ITCS.2022.79. URL https://drops-dev.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2022.79.

Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss Minimization Through the Lens Of Outcome Indistinguishability. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 60:1–60:20, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-263-1. doi: 10.4230/LIPIcs.ITCS.2023.60. URL https://drops-dev.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2023.60.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Nika Haghtalab, Mingda Qiao, Kunhe Yang, and Eric Zhao. Truthfulness of calibration measures. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Lunjia Hu and Yifan Wu. Predict to minimize swap regret for all payoff-bounded tasks. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 244–263. IEEE, 2024.

Lunjia Hu, Arun Jambulapati, Kevin Tian, and Chutong Yang. Testing calibration in nearly-linear time. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Sham M. Kakade and Dean P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2007.04.017. URL https://www.sciencedirect.com/science/article/pii/S0022000007000633. Learning Theory 2004.

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5143–5145. PMLR, 2023.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2805–2814. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kumar18a.html.

Nicolas S Lambert. Elicitation and evaluation of statistical forecasts. *Preprint*, 2011.

Yingkai Li, Jason D Hartline, Liren Shan, and Yifan Wu. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 988–989, 2022.

Yuxuan Lu, Yifan Wu, Jason Hartline, and Lunjia Hu. Truthful calibration errors for multi-class prediction. *arXiv preprint arXiv:2510.06388*, 2025.

John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42(9):654, 1956.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.

Princewill Okoroafor, Robert Kleinberg, and Michael P Kim. Near-optimal algorithms for omniprediction. *arXiv preprint arXiv:2501.17205*, 2025.

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 Eric Zhao. Truthfulness of decision-theoretic calibration measures. *arXiv preprint arXiv:2503.02384*, 2025.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL https://doi.org/10.1137/1116025.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Robert L Winkler, Javier Munoz, José L Cervera, José M Bernardo, Gail Blattenberger, Joseph B Kadane, Dennis V Lindley, Allan H Murphy, Robert M Oliver, and David Ríos-Insua. Scoring rules and the evaluation of probabilities. *Test*, 5:1–60, 1996.

In Appendix D, we show that Averaged Two-Bin Calibration Error (ATB) is optimally valid for smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL), implying a linear-time calibration tester for SMCAL and DISTCAL.

## Appendix A. Preliminaries

Throughout the paper, we use $D$ to denote a joint distribution of $(x, y)$ pairs, where $x \in X$ represents an individual in a domain $X$, and $y \in \{0, 1\}$ is the corresponding state (a.k.a. outcome or label). A predictor $r : X \to [0, 1]$ reports a prediction $r(x) \in [0, 1]$ for each individual $x \in X$.

We present useful definitions and preliminary theorems for our paper. Appendix A.1 introduces existing calibration measures. Appendix A.2 defines the completeness and soundness of a calibration measure as well as consistency of its estimator. Appendix A.3 formalizes truthfulness of an error measure. Appendix A.4 introduces plug-in calibration tests, preparing for the linear-time calibration tester result.

### A.1. Calibration

We present the formal definitions of a few important calibration error metrics in the literature. We start with the definition of calibration:

**Definition 4 (Calibration)** *A predictor $r : X \to [0, 1]$ is calibrated on an underlying distribution $D$ of $(x, y) \in X \times \{0, 1\}$ if $\mathbb{E}_D[y|r(x)] = r(x)$ holds almost surely.*

An important property of the definition of calibration is that it only depends on the distribution of the prediction-state pair $(r(x), y) \in [0, 1] \times \{0, 1\}$. That is, we can determine whether a predictor $r$ is calibrated on a distribution $D$ just based on the distribution of $(r(x), y)$, without having to know the full joint distribution of $(x, r(x), y)$. Thus, using a random variable $v$ to represent the prediction value $r(x)$, we can define calibration simply given a distribution $J$ of $(v, y) \in [0, 1] \times \{0, 1\}$:

**Definition 5 (Calibration of prediction-state distributions)** *We say a distribution $J$ of $(v, y) \in [0, 1] \times \{0, 1\}$ is calibrated if $\mathbb{E}_J[y|v] = v$ holds almost surely.*

For a distribution $D$ of $(x, y) \in X \times \{0, 1\}$ and a predictor $r : X \to [0, 1]$, we use $J_{D,r}$ to denote the joint distribution of $(r(x), y)$. With that, $r$ is calibrated on $D$ if and only if $J_{D,r}$ is calibrated as in Definition 5.

A calibration measure $\textsc{Cal}_D(r) \in \mathbb{R}_{\geq 0}$ evaluates the deviation of a predictor $r$ from being perfectly calibrated on a distribution $D$. Naturally, we define a calibration measure $\textsc{Cal}(J)$ first for general prediction-state distributions $J$ of $(v, y) \in [0, 1] \times \{0, 1\}$, and then define

$$\textsc{Cal}_D(r) := \textsc{Cal}(J_{D,r}).$$

The most well-known calibration measure is the *Expected Calibration Error (ECE)*:

**Definition 6 (Expected Calibration Error (ECE), Foster and Vohra 1997)** *Let $J$ be a distribution of $(v, y) \in [0, 1] \times \{0, 1\}$, and let random variable $\widehat{v} := \mathbb{E}_J[y|v]$ be the conditional expectation of the state $y$ given the prediction value $v$. The Expected Calibration Error (ECE) is defined as*

$$\text{ECE}(J) := \mathbb{E}\,|v - \widehat{v}|.$$

*Correspondingly, for a distribution $D$ of $(x, y) \sim X \times \{0, 1\}$ and a predictor $r : X \to [0, 1]$, defining $\widehat{r}(x) := \mathbb{E}_D[y|r(x)]$, we have*

$$\text{ECE}_D(r) := \text{ECE}(J_{D,r}) = \mathbb{E}_D\,|r(x) - \widehat{r}(x)|.$$

*More generally, for every $\alpha \geq 1$, we define $\ell_\alpha$ Expected Calibration Error (ECE) as follows:*

$$\ell_\alpha\text{-ECE}(J) := \mathbb{E}[|v - \widehat{v}|^\alpha], \quad \ell_\alpha\text{-ECE}_D(r) := \mathbb{E}_D[|r(x) - \widehat{r}(x)|^\alpha].$$

A downside of the Expected Calibration Error (ECE) is its discontinuity: slight changes in the predictions $r(x)$ can cause significant changes to the ECE value. This motivated Błasiok et al. (2023) to introduce a continuous calibration error metric, termed the *distance to calibration*. It measures the earthmover distance from the prediction-state distribution $(v, y)$ to a calibrated distribution $(u, y)$.

**Definition 7 ((Lower) Distance to Calibration (DISTCAL), Błasiok et al. 2023)** *Let $J$ be a distribution of $(v, y) \in [0,1] \times \{0,1\}$. Consider a joint distribution (i.e. coupling) $\Pi$ of $(u, v, y) \in [0,1] \times [0,1] \times \{0,1\}$, where $(v, y)$ is distributed according to $J$, and the distribution of $(u, y)$ is calibrated as in Definition 5. The lower distance to calibration (DISTCAL) is defined as the following infimum over all such couplings $\Pi$:*

$$\underline{\text{DISTCAL}}(J) := \inf_{\Pi} \mathbb{E}_{\Pi} |u - v|.$$

*Correspondingly, given a distribution $D$ of $(x, y) \sim X \times \{0,1\}$ and a predictor $r : X \to [0,1]$, we define $\underline{\text{DISTCAL}}_D(r) := \underline{\text{DISTCAL}}(J_{D,r})$.*

One might imagine a different definition of the distance to calibration as the minimum $L_1$ distance $\mathbb{E}_D |r(x) - r'(x)|$ from the given predictor $r$ to a calibrated predictor $r'$. Indeed, this distance to calibration notion (denoted DISTCAL) is the first definition of the distance to calibration introduced by Błasiok et al. (2023). However, as shown by Błasiok et al. (2023), this definition is different from the DISTCAL in Definition 7 and has the disadvantage of depending on the full joint distribution of $(x, r(x), y)$, not just the prediction-state distribution of $(r(x), y)$. To address this disadvantage, Błasiok et al. (2023) introduced the DISTCAL in Definition 7 and termed it the *lower* distance to calibration. They also showed that the two definitions are quadratically related:

$$\frac{1}{16} \text{DISTCAL}_D(r)^2 \le \underline{\text{DISTCAL}}_D(r) \le \text{DISTCAL}_D(r).$$

We will focus on the lower distance to calibration in Definition 7 throughout the paper and will often drop the word "lower" for brevity.

Another important continuous calibration measure is the *smooth calibration error* introduced by Kakade and Foster (2008) (originally termed *weak calibration*). As shown by Błasiok et al. (2023), the smooth calibration error (SMCAL) gives a constant factor approximation to lower distance to calibration (DISTCAL) (see Proposition 9 below).

**Definition 8 (Smooth Calibration Error (SMCAL), (Kakade and Foster, 2008))** *Let $W_1$ be the family of 1-Lipschitz functions $w : [0,1] \to [-1,1]$. For any distribution $J$ of $(v, y) \in [0,1] \times \{0,1\}$, the smooth calibration error is defined as*

$$\text{SMCAL}(J) := \sup_{w \in W_1} \mathbb{E}_J[(v - y)w(v)]. \tag{15}$$

*Correspondingly, for a distribution $D$ of $(x, y) \sim X \times \{0,1\}$ and a predictor $r : X \to [0,1]$, we have*

$$\text{SMCAL}_D(r) := \text{SMCAL}(J_{D,r}) = \sup_{w \in W_1} \mathbb{E}_J[(r(x) - y)w(r(x))].$$

Without the Lipschitzness constraint on $w$, the smooth calibration error would become the Expected Calibration Error (ECE) (Definition 6), where the supremum in (15) is achieved by

$$w(v) = \begin{cases} 1, & \text{if } \widehat{v} > v; \\ -1, & \text{otherwise.} \end{cases}$$

The following proposition shows that DISTCAL and SMCAL are constant factor approximations of each other:

**Proposition 9 (Błasiok et al. 2023)** *For any distribution $J$ of $(v, y) \in [0,1] \times \{0,1\}$,*

$$\frac{1}{2} \underline{\text{DISTCAL}}(J) \le \text{SMCAL}(J) \le 2\, \underline{\text{DISTCAL}}(J).$$

### A.2. Completeness, Soundness, and Estimation

A basic property shared by all the calibration measures in Appendix A.1 is that they are all minimized when the predictor is calibrated, with the minimum value being zero:

**Claim 10** *For* CAL *equal to Expected Calibration Error (*ECE*), lower distance to calibration (*DISTCAL*), or smooth calibration error (*SMCAL*), we have* $\text{CAL}(J) \geq 0$ *for any distribution* $J$ *of* $(v, y) \in [0, 1] \times \{0, 1\}$. *Moreover,*

$$\text{CAL}(J) = 0 \iff J \text{ is calibrated (Definition 5)}.$$

The claim above is a distributional statement: if we know the prediction-state distribution $J$ exactly, then $\text{CAL}(J)$ detects calibration by whether it is zero. We use completeness and soundness for this population target.

**Definition 11 (Completeness)** *We say a calibration measure* CAL *is complete if* $\text{CAL}(J) = 0$ *for every calibrated distribution* $J$ *of prediction-state pairs* $(v, y) \in [0, 1] \times \{0, 1\}$.

**Definition 12 (Soundness)** *We say a calibration measure* CAL *is sound if* $\text{CAL}(J) > 0$ *for every mis-calibrated distribution* $J$ *of prediction-state pairs* $(v, y) \in [0, 1] \times \{0, 1\}$.

Together, completeness and soundness say exactly that $\text{CAL}(J) = 0$ if and only if $J$ is calibrated. In practice, however, we rarely get access to the full distribution $J$ and instead evaluate calibration from an independent and identically distributed (i.i.d.) sample. We therefore separate the population calibration measure from the estimator used to approximate it.

**Definition 13 (Consistent estimator)** *Let* $\widehat{\text{CAL}}_T$ *be an estimator that maps a sample to a nonnegative real number. We say* $\{\widehat{\text{CAL}}_T\}_{T\geq 1}$ *is consistent for a calibration measure* CAL *if, for every distribution* $J$ *of prediction-state pairs,*

$$\lim_{T\to\infty} \mathbb{E}_{S\sim J^T}\left[\left|\widehat{\text{CAL}}_T(S) - \text{CAL}(J)\right|\right] = 0.$$

*The plug-in estimator of* CAL *is the estimator* $\widehat{\text{CAL}}_T^{\text{plug}}(S) := \text{CAL}(J_S)$, *where* $J_S$ *is the empirical distribution over* $S$.

**Proposition 14** *If* CAL *is complete and sound and* $\{\widehat{\text{CAL}}_T\}_{T\geq 1}$ *is consistent for* CAL, *then the estimator has vanishing expected error on calibrated distributions and non-vanishing expected error on every fixed mis-calibrated distribution:*

$$\lim_{T\to\infty} \mathbb{E}_{S\sim J^T}[\widehat{\text{CAL}}_T(S)] = 0 \quad \textit{for calibrated } J,$$

*and*

$$\liminf_{T\to\infty} \mathbb{E}_{S\sim J^T}[\widehat{\text{CAL}}_T(S)] > 0 \quad \textit{for mis-calibrated } J.$$

**Proof** If $J$ is calibrated, completeness gives $\text{CAL}(J) = 0$, so the first limit follows from the consistency of $\widehat{\text{CAL}}_T$. If $J$ is mis-calibrated, soundness gives $\text{CAL}(J) > 0$, and

$$\mathbb{E}_{S\sim J^T}[\widehat{\text{CAL}}_T(S)] \geq \text{CAL}(J) - \mathbb{E}_{S\sim J^T}\left[\left|\widehat{\text{CAL}}_T(S) - \text{CAL}(J)\right|\right].$$

Taking the limit inferior proves the second claim. ■

**Relation to the previous empirical definitions.** In earlier drafts, completeness and soundness were defined directly for the plug-in estimate $\text{CAL}(J_S)$. In that terminology, CAL was called complete if

$$\lim_{T\to\infty} \mathbb{E}_{S\sim J^T}[\text{CAL}(J_S)] = 0$$

for every calibrated $J$, and sound if

$$\liminf_{T\to\infty} \mathbb{E}_{S\sim J^T}[\text{CAL}(J_S)] > 0$$

for every fixed mis-calibrated $J$. This is exactly the conclusion of Proposition 14 when the plug-in estimator $\text{CAL}(J_S)$ is consistent for a complete and sound distributional measure CAL.

It can be inferred from the work of Błasiok et al. (2023) that smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL) are both complete and sound. Moreover, while Expected Calibration Error (ECE) satisfies Claim 10 and is therefore complete and sound as a distributional measure, its plug-in estimator is not consistent. To see this, consider the distribution $J$ of prediction-state pairs $(v, y) \in [0, 1] \times \{0, 1\}$, where $v$ is drawn uniformly from $[1/3, 2/3]$, and conditioned on $v$, $y$ is drawn from the Bernoulli distribution with mean $v$. Clearly, $J$ is calibrated and $\text{ECE}(J) = 0$. However, on a finite sample $S = \{(v_1, y_1), \ldots, (v_T, y_T)\}$ of independent and identically distributed (i.i.d.) draws from $J$, it holds almost surely that all the $v_t$'s are distinct, in which case $\text{ECE}(J_S) \geq 1/3$ (see Example 1).

Due to the inconsistency of the plug-in estimator for Expected Calibration Error (ECE), in machine learning practice, the binned Expected Calibration Error (BINECE) is widely adopted as an empirical method for estimating ECE (Guo et al., 2017; Minderer et al., 2021). While we do not need this notion to state our main results, we include its definition here for completeness:

**Definition 15 (Binned Expected Calibration Error (Binned ECE))** *Let $J$ be a distribution of $(v, y) \in [0, 1] \times \{0, 1\}$. Given $\alpha \geq 1$ and a partition $\mathcal{I} = \{I_i\}_{i\in[k]}$ of the prediction space $[0, 1]$, the $\ell_\alpha$-binned* ECE *is defined as*

$$\ell_\alpha\text{-BINECE}(J) := \sum_{i\in[k]} \Pr_J[v \in I_i] \cdot \left|\mathbb{E}_J[v - y | v \in I_i]\right|^\alpha.$$

*We take the contribution of a zero-mass bin to be zero.*

We can estimate $\ell_\alpha$-BINECE$(J)$ using a sample $S = \{(v_t, y_t)\}_{t\in[T]}$ of $T$ independent and identically distributed (i.i.d.) points drawn from $J$. Specifically, letting $J_S$ denote the (empirical) uniform distribution over $S$, we can use $\ell_\alpha$-BINECE$(J_S)$ as a good estimate for $\ell_\alpha$-BINECE$(J)$ when the sample size $T$ is sufficiently large relative to $k$ (the number of bins). In practice, the number $k$ of bins can be selected according to the sample size $T$, e.g. $k = T^{\frac{1}{3}}$, to obtain a consistent estimator and hence recover the empirical completeness and soundness guarantees above.

**Remark 16 (Comparison to Haghtalab et al. 2024)** *The empirical formulation above follows the same idea as Haghtalab et al. (2024), while our empirical soundness requirement is strictly stronger. There exists an error metric that is not reasonably sound but satisfies the completeness and soundness in Haghtalab et al. (2024).*

*The soundness definition in Haghtalab et al. (2024) requires that for any empirical distribution $D_T$ over $T$ samples,*

- *if* $r_t = 1 - y_t$ *for all* $t$*, then* $\lim_{T\to\infty} \textsc{Cal}_T(r) = \Omega(1)$*;*
- *if each state* $y \sim$ *Ber*$(\alpha)$ *is an independent Bernoulli variable with the same mean, then* $\lim_{T\to\infty} \textsc{Cal}_T(r) = \Omega(1)$ *for any non-truthful constant report* $r \neq \beta$*.*

*We see that the error* $\textsc{Cal} = (\mathbb{E}[r] - \mathbb{E}[y])^2 + \mathbb{E}[\mathbb{I}[r \in \{0,1\}, y \neq r]]$ *satisfies the requirements above. However, for predictions not in* $\{0,1\}$*, the error metric only evaluates the unconditional bias in predictions, which is far from a calibration error metric.*

### A.3. Truthfulness

A truthful error metric incentivizes a strategizing predictor to report the true distribution to minimize expected error on a finite sample. Definition 17 defines ex-ante truthfulness, where a predictor output is assumed to be a function of the feature space.

**Definition 17 (Ex-Ante Truthfulness)** *We say a calibration measure* $\textsc{Cal}$ *is ex-ante truthful if the following holds. Let* $D$ *be an arbitrary joint distribution of* $(x, y) \in X \times \{0,1\}$ *and let* $p : X \to [0,1]$ *be the ground-truth predictor* $p(x) = \mathbb{E}_D[y|x]$*. Let* $S = \{(x_t, y_t)\}_{t\in[T]}$ *be a sample of* $T$ *independent and identically distributed (i.i.d.) points drawn from* $D$*, and let* $D_S$ *denote the (empirical) uniform distribution over* $S$*. Then*

$$\mathbb{E}_S[\textsc{Cal}_{D_S}(p)] \leq \mathbb{E}_S[\textsc{Cal}_{D_S}(r)] \quad \textit{for every predictor } r : X \to [0,1].$$

In this paper, we study a strictly stronger notion: interim truthfulness. In the interim stage, the true distributions of $T$ samples are realized, and the predictor is allowed to deviate and report any prediction sequence, not necessarily a function of the feature space. We first extend our definition of calibration errors to this setting, where we evaluate the calibration error of a reported prediction sequence $\boldsymbol{r} = (r_1, \ldots, r_T)$ for the $T$ individuals w.r.t. a ground-truth probability sequence $\boldsymbol{p} = (p_1, \ldots, p_T)$. We will refer to both definitions as truthfulness when it is clear from the context.

**Definition 18 (Induced calibration error on prediction sequences)** *Given a calibration measure* $\textsc{Cal}(J)$ *defined on prediction-state distributions* $J$ *over* $[0,1] \times \{0,1\}$*, we define an induced calibration measure* $\textsc{Cal}(\boldsymbol{r}, \boldsymbol{p})$ *as follows, where* $\boldsymbol{r} = (r_1, \ldots, r_T) \in [0,1]^T$ *is a sequence of predictions and* $\boldsymbol{p} = (p_1, \ldots, p_T) \in [0,1]^T$ *is a sequence of ground-truth probabilities. Let* $J_{\boldsymbol{r},\boldsymbol{p}}$ *be the distribution of* $(r_t, y) \in [0,1] \times \{0,1\}$ *where* $t$ *is drawn uniformly from* $[T]$*, and* $y \in \{0,1\}$ *is drawn from the Bernoulli distribution with mean* $p_t$*. We define*

$$\textsc{Cal}(\boldsymbol{r}, \boldsymbol{p}) := \textsc{Cal}(J_{\boldsymbol{r},\boldsymbol{p}}).$$

For example, according to Definition 18, we can explicitly calculate Expected Calibration Error (ECE) and smooth calibration error (SMCAL) as follows. Recall that for $v \in \{r_1, \ldots, r_T\}$, we define

$$\widehat{v} := \mathbb{E}_{(v,y)\sim J_{\boldsymbol{r},\boldsymbol{p}}}[y|v] = \frac{\sum_{t\in[T]} p_t \mathbb{I}[r_t = v]}{\sum_{t\in[T]} \mathbb{I}[r_t = v]}. \tag{16}$$

We have

$$\text{Expected Calibration Error (ECE):} \quad \mathrm{ECE}(\boldsymbol{r},\boldsymbol{p}) = \mathrm{ECE}(J_{\boldsymbol{r},\boldsymbol{p}}) = \mathbb{E}_{(v,y)\sim J_{\boldsymbol{r},\boldsymbol{p}}}\left[\left|v-\widehat{v}\right|\right]$$
$$= \frac{1}{T}\sum_{v}\sum_{t\in[T]}\mathbb{I}\left[r_t = v\right]\left|v-\widehat{v}\right|$$
($v$ ranges over all values that appear at least once in the set $\{r_1,\dots r_t\}$)
$$= \frac{1}{T}\sum_{v}\left|(v-\widehat{v})\sum_{t\in[T]}\mathbb{I}\left[r_t=v\right]\right|$$
$$= \frac{1}{T}\sum_{v}\left|\sum_{t\in[T]}(r_t-p_t)\mathbb{I}\left[r_t=v\right]\right|.$$
(by (16) and $v\mathbb{I}\left[r_t=v\right] = r_t\mathbb{I}\left[r_t=v\right]$)

Similarly for smooth calibration error (SMCAL):

$$\mathrm{SMCAL}(\boldsymbol{r},\boldsymbol{p}) = \sup_{w\in W_1}\frac{1}{T}\sum_{t=1}^{T}(r_t-p_t)w(r_t). \qquad (W_1 \text{ is the same as in Definition 8})$$

We now define the notion of truthfulness for the calibration errors from Definition 18 on length-$T$ sequences. We note that this definition is akin to the definition of properness in the literature of proper scoring rules (McCarthy, 1956; Savage, 1971).

**Definition 19 (Interim Truthfulness)** *We say a calibration measure* CAL *is interim truthful if the following holds for any $T\in\mathbb{Z}_{>0}$. Let $\boldsymbol{p} := (p_1,\dots,p_T)\in[0,1]^T$ be an arbitrary sequence of ground-truth predictions. Let $\boldsymbol{y} = (y_1,\dots,y_T)$ denote the randomly realized states, where each $y_t\in\{0,1\}$ is drawn independently from the Bernoulli distribution with mean $p_t$ (denoted $\boldsymbol{y}\sim\boldsymbol{p}$). Then*

$$\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\mathrm{CAL}(\boldsymbol{p},\boldsymbol{y})] \le \mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\mathrm{CAL}(\boldsymbol{r},\boldsymbol{y})] \quad \textit{for any } \boldsymbol{r} = (r_1,\dots,r_T)\in[0,1]^T.$$

**Claim 20 (Interim truthfulness implies ex-ante truthfulness)** *Let $\mathrm{CAL}(\boldsymbol{r},\boldsymbol{p})$ be a calibration measure induced by $\mathrm{CAL}(J)$ (Definition 18). If $\mathrm{CAL}(\boldsymbol{r},\boldsymbol{p})$ is interim truthful, then $\mathrm{CAL}(J)$ is ex-ante truthful.*

**Proof** As in Definition 17, consider a sample $S = \{(x_1,y_1),\dots,(x_T,y_T)\}$ of independent and identically distributed (i.i.d.) points from a distribution $D$ over $X\times\{0,1\}$, and let $r: X\to[0,1]$ be a predictor. Define $\boldsymbol{r} := (r(x_1),\dots,r(x_T))$ and $\boldsymbol{y} := (y_1,\dots,y_T)$. Now $J_{D_S,r}$ and $J_{\boldsymbol{r},\boldsymbol{y}}$ are both equal to the distribution of $(r(x_t),y_t)$ for uniform $t\in[T]$. Therefore,

$$\mathrm{CAL}_{D_S}(r) = \mathrm{CAL}(J_{D_S,r}) = \mathrm{CAL}(J_{\boldsymbol{r},\boldsymbol{y}}) = \mathrm{CAL}(\boldsymbol{r},\boldsymbol{y}). \tag{17}$$

As in Definition 17, define $p(x_t) := \mathbb{E}_D[y|x=x_t]\in[0,1]$ for $t=1,\dots,T$. Conditioned on $x_1,\dots,x_T$, each $y_t$ is distributed independently from the Bernoulli distribution with mean $p(x_t)$. That is, we have $\boldsymbol{y}\sim\boldsymbol{p}$ as in Definition 19, where $\boldsymbol{p} := (p(x_1),\dots,p(x_T))$. Therefore, by (17),

$$\mathbb{E}_S[\mathrm{CAL}_{D_S}(r)|x_1,\dots,x_T] = \mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\mathrm{CAL}(\boldsymbol{r},\boldsymbol{y})], \tag{18}$$
$$\mathbb{E}_S[\mathrm{CAL}_{D_S}(p)|x_1,\dots,x_T] = \mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\mathrm{CAL}(\boldsymbol{p},\boldsymbol{y})]. \tag{19}$$

Assuming interim truthfulness, we know that the quantity in (18) is no smaller than the quantity in (19). Taking the expectation over $x_1, \ldots, x_T$ proves the desired ex-ante truthfulness. ■

### A.4. Calibration Test and Validity

Completeness and soundness (Definitions 11 and 12) ensure that the population calibration measure CAL distinguishes calibrated predictors from mis-calibrated ones. Together with a consistent plug-in estimator (Definition 13), this distinction can be recovered from samples when the sample size $T$ is large enough. Intuitively, we should expect the distinguishing power to grow as a function of $T$. We characterize this quantitative dependence on $T$ below. We first define plug-in calibration tests that aim at accepting calibrated predictors while rejecting mis-calibrated ones, based on a sample of size $T$.

**Definition 21 (Plug-in Calibration Test)** *Consider the following calibration test using a calibration measure* CAL. *Let $J$ be an arbitrary distribution of prediction-state pairs $(v, y) \in [0,1] \times \{0,1\}$. The test first draws $T$ independent and identically distributed (i.i.d.) points from $J$ to form a sample $S = \{(v_t, y_t)\}_{t\in[T]}$, and then computes the calibration error $\textsc{Cal}(J_S)$ on the uniform distribution $J_S$ over $S$. The test outputs "accept" if the calibration error does not exceed a threshold $\beta$. That is, the acceptance probability of this test is*

$$\mathsf{accP}^{\textsc{Cal}}(J; T, \beta) := \Pr_{S\sim J^T}[\textsc{Cal}(J_S) \le \beta].$$

We define the validity of a calibration measure CAL given a reference calibration measure REF that is often chosen to be complete and sound.

**Definition 22 (Validity)** *Let $\{\gamma_T\}$ be an infinite sequence of real numbers indexed by $T = 1, 2, \ldots$. We say a calibration measure* CAL *is $\{\gamma_T\}$-valid w.r.t. a reference calibration measure* REF *if there exist thresholds $\beta_1, \beta_2, \ldots \in \mathbb{R}$ such that*

$$\liminf_{T\to\infty} \left( \inf_{J\,:\,\textit{calibrated}} \mathsf{accP}^{\textsc{Cal}}(J; T, \beta_T) - \sup_{J\,:\,\textsc{Ref}(J)\ge\gamma_T} \mathsf{accP}^{\textsc{Cal}}(J; T, \beta_T) \right) > 0.$$

*That is, there is a non-vanishing gap between the acceptance probability when $J$ is calibrated, and the acceptance probability when $J$ is mis-calibrated with error at least $\gamma_T$ in the reference measure* REF.

In the definition above, one should typically think of $\gamma_T$ as a decreasing function of $T$, which indicates the stronger distinguishing power as $T$ grows. Moreover, the faster $\gamma_T$ decreases, the stronger is the distinguishing power of a $\{\gamma_T\}$-valid calibration error for large $T$.

## Appendix B. Truthful Family: Unnormalized Binned Squared Errors

In this section, we present a general family of truthful error metrics, which we term *Unnormalized Binned Squared Errors (UBSEs)*. As will become clear, Averaged Two-Bin Calibration Error (ATB) is a special case of UBSEs, so its truthfulness is an immediate consequence of the truthfulness of UBSEs.

**Definition 23 (Unnormalized Binned Squared Errors)** *Consider an error metric* $\textsc{Cal}(\boldsymbol{r}, \boldsymbol{p})$ *taking as input a report vector* $\boldsymbol{r} = (r_1, \ldots, r_T) \in [0,1]^T$ *and a ground-truth vector* $\boldsymbol{p} = (p_1, \ldots, p_T) \in [0,1]^T$. *We say* CAL *is an* unnormalized binned squared error *(UBSE) if it can be calculated as follows:*

1. *Partition the indices* $[T]$ *into* $k$ *disjoint bins:* $[T] = B_1 \cup \cdots \cup B_k$. *Importantly, we allow the partition (including the choice of* $k$*) to be randomized, and we allow it to depend on the report vector* $\boldsymbol{r}$ *(but not on* $\boldsymbol{p}$*).*

2. *Compute the bias* $\Delta_i$ *in each bin* $B_i$*:*

$$\Delta_i := \frac{1}{T} \sum_{t \in B_i} (r_t - p_t). \tag{20}$$

3. *Output the sum of the squared biases:* $\textsc{Cal}(\boldsymbol{r}, \boldsymbol{p}) := \mathbb{E}_{\mathcal{B}}[\sum_{i=1}^k \Delta_i^2]$*, where the expectation is over the randomness of the partition* $\mathcal{B} = (B_1, \ldots, B_k)$.

The above definition is very similar to the definition of binned $\ell_2$ Expected Calibration Error (ECE), $\ell_2$-ECE, but there is a crucial difference. When defining binned $\ell_2$-ECE for a fixed partition $\mathcal{B} = (B_1, \ldots, B_k)$, the bias in each bin is first *normalized by the bin size* $|B_i|$:

$$\widetilde{\Delta}_i = \frac{1}{|B_i|} \sum_{t \in B_i} (r_t - p_t),$$

and then squared and summed with *weights* $|B_i|/T$:

$$\ell_2\text{-ECE}(\boldsymbol{r}, \boldsymbol{p}) = \sum_{i=1}^k \frac{|B_i|}{T} \widetilde{\Delta}_i^2 = \sum_{i=1}^k \frac{1}{|B_i| T} \left( \sum_{t \in B_i} (r_t - p_t) \right)^2.$$

In contrast, Definition 23 takes the *unweighted* sum of the *unnormalized* squared biases $\Delta_i^2$:

$$\textsc{Cal}(\boldsymbol{r}, \boldsymbol{p}) = \mathbb{E}_{\mathcal{B}} \left[ \sum_{i=1}^k \Delta_i^2 \right] = \mathbb{E}_{\mathcal{B}} \left[ \sum_{i=1}^k \frac{1}{T^2} \left( \sum_{t \in B_i} (r_t - p_t) \right)^2 \right].$$

### B.1. Interim Truthfulness

Unnormalized Binned Squared Errors (UBSEs) are interim truthful (whereas the binned $\ell_2$-ECE is not, with the small difference above):

**Theorem 24** *Any UBSE error metric* CAL *is interim truthful (Definition 19).*

In fact, we prove a stronger result in Lemma 25, showing that the expected empirical Unnormalized Binned Squared Error (UBSE) decomposes into the UBSE on the true probabilities $\boldsymbol{p}$ plus a variance term independent of $\boldsymbol{r}$.

**Lemma 25 (Error Decomposition)** *Let* CAL *be an arbitrary Unnormalized Binned Squared Error (UBSE). For any report sequence* $\boldsymbol{r} = (r_1, \ldots, r_T) \in [0,1]^T$ *and any ground-truth vector* $\boldsymbol{p} = (p_1, \ldots, p_T) \in [0,1]^T$,

$$\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\textsc{Cal}(\boldsymbol{r}, \boldsymbol{y})] = \textsc{Cal}(\boldsymbol{r}, \boldsymbol{p}) + \frac{1}{T^2}\sum_{t=1}^{T} p_t(1-p_t).$$

*Here* $\boldsymbol{y} = (y_1, \ldots, y_T) \in \{0,1\}^T$ *is drawn such that each* $y_t$ *independently follows the Bernoulli distribution with mean* $p_t$ *(as in Definition 19).*

We first prove Theorem 24 using Lemma 25, and then prove Lemma 25.

**Proof** [Proof of Theorem 24] For any $\boldsymbol{r}, \boldsymbol{p} \in [0,1]^T$, by Lemma 25,

$$\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\textsc{Cal}(\boldsymbol{r}, \boldsymbol{y})] - \mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\textsc{Cal}(\boldsymbol{p}, \boldsymbol{y})] = \textsc{Cal}(\boldsymbol{r}, \boldsymbol{p}) - \textsc{Cal}(\boldsymbol{p}, \boldsymbol{p}). \tag{21}$$

Clearly, we have $\textsc{Cal}(\boldsymbol{r}, \boldsymbol{p}) \geq 0$ and $\textsc{Cal}(\boldsymbol{p}, \boldsymbol{p}) = 0$. Therefore, the quantity in (21) is non-negative, which means that CAL is interim truthful. ■

**Proof** [Proof of Lemma 25] For a partition $\mathcal{B} = (B_1, \ldots, B_k)$ of $[T]$ as in Definition 23, we define

$$\Delta_i := \frac{1}{T}\sum_{t\in B_i}(r_t - y_t).$$
$$\widehat{\Delta}_i := \frac{1}{T}\sum_{t\in B_i}(r_t - p_t).$$

We have

$$\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\textsc{Cal}(\boldsymbol{r}, \boldsymbol{y})] = \mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}\left[\mathbb{E}_{\mathcal{B}}\left[\sum_{i=1}^{k}\Delta_i^2\right]\right] = \mathbb{E}_{\mathcal{B}}\left[\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}\left[\sum_{i=1}^{k}\Delta_i^2\right]\right], \tag{22}$$

$$\textsc{Cal}(\boldsymbol{r}, \boldsymbol{p}) = \mathbb{E}_{\mathcal{B}}\left[\sum_{i=1}^{k}\widehat{\Delta}_i^2\right]. \tag{23}$$

In (22), we used the fact that the distribution of $\mathcal{B}$ depends only on $\boldsymbol{r}$ and not on $\boldsymbol{y}$. For the same reason, the two distributions of $\mathcal{B}$ in (22) and (23) are the same. Therefore, to prove the lemma, it suffices to show that for any fixed partition $\mathcal{B}$,

$$\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}\left[\sum_{i=1}^{k}\Delta_i^2\right] = \sum_{i=1}^{k}\widehat{\Delta}_i^2 + \frac{1}{T^2}\sum_{t=1}^{T}p_t(1-p_t). \tag{24}$$

For every $i = 1, \ldots, k$, we have

$$\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\Delta_i^2] = \mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\Delta_i]^2 + \textsc{Var}_{\boldsymbol{y}\sim\boldsymbol{p}}[\Delta_i], \tag{25}$$

where

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\Delta_i] &= \widehat{\Delta}_i, \\
\textsc{Var}_{\boldsymbol{y}\sim\boldsymbol{p}}[\Delta_i] &= \textsc{Var}_{\boldsymbol{y}\sim\boldsymbol{p}}\left[\frac{1}{T}\sum_{t\in B_i}(r_t - y_t)\right] \\
&= \frac{1}{T^2}\textsc{Var}_{\boldsymbol{y}\sim\boldsymbol{p}}\left[\sum_{t\in B_i} y_t\right] \\
&= \frac{1}{T^2}\sum_{t\in B_i}\textsc{Var}_{\boldsymbol{y}\sim\boldsymbol{p}}[y_t] && \text{(the } y_t\text{'s are distributed independently)} \\
&= \frac{1}{T^2}\sum_{t\in B_i} p_t(1-p_t).
\end{aligned}$$

Plugging these into (25), we have

$$\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\Delta_i^2] = \widehat{\Delta}_i^2 + \frac{1}{T^2}\sum_{t\in B_i} p_t(1-p_t).$$

Summing up over $i = 1, \ldots, k$ proves (24). ■

We remark that in addition to being truthful, Unnormalized Binned Squared Errors (UBSEs) have a vanishing empirical-error guarantee for calibrated finite prediction sequences: by Lemma 25, the expected error of calibrated finite prediction sequences is

$$\frac{1}{T^2}\sum_{t=1}^{T} p_t(1-p_t) \le \frac{1}{4T} = O(1/T),$$

which vanishes as $T \to \infty$.

**Example 2 (Quantile-Binned $\ell_2$ Expected Calibration Error (ECE) is truthful)** *As a special case of an Unnormalized Binned Squared Error (UBSE), the quantile-binned $\ell_2$ Expected Calibration Error (ECE), $\ell_2$-*ECE*, is truthful and has the same empirical completeness guarantee. Choosing the number of bins properly as a growing function of $T$, it is also empirically sound in the sample-access sense discussed in Appendix A.2. It is defined as follows:*

*For any report sequence $\boldsymbol{r} = (r_1, \ldots, r_T)$ and any vector of realized state $\boldsymbol{y} = (y_1, \ldots, y_T)$,*

- *sort the predictions in increasing order with $\boldsymbol{r}_1 \le \cdots \le \boldsymbol{r}_T$, break ties uniformly at random.*
- *Partition predictions into $k = T^{1/3}$ bins by quantile. Each bin has $\frac{T}{k}$ predictions.*
- *Given the partition above, output the Unnormalized Binned Squared Error* $\textsc{Cal}(\boldsymbol{r}, \boldsymbol{y})$.

### B.2. Strict Ex-Ante Truthfulness

In the ex-ante stage before the ground-truth probabilities for each sample are drawn, an Unnormalized Binned Squared Error (UBSE) is strictly truthful, i.e., the unique minimizer to the expected error is the ground truth predictor $p(x) := \mathbb{E}[y|x]$.

**Theorem 26** *Let $X$ be an arbitrary non-empty domain and let $D$ be an arbitrary distribution of $(x, y) \in X \times \{0, 1\}$. Let $p : X \to [0, 1]$ be the ground-truth predictor $p(x) := \mathbb{E}[y|x]$ and let $r : X \to [0, 1]$ be an arbitrary predictor. Let* CAL *be an arbitrary Unnormalized Binned Squared Error (UBSE) and $T$ be an arbitrary positive integer. For a sample $S = ((x_t, y_t))_{t\in[T]}$ of $T$ independent and identically distributed (i.i.d.) examples drawn from $D$, let $D_S$ denote the uniform distribution on $S$. Suppose*

$$\mathbb{E}_S[\textsc{Cal}_{D_S}(r)] \le \mathbb{E}_S[\textsc{Cal}_{D_S}(p)].$$

*Then $r(x) = p(x)$ holds almost surely over $(x, y) \sim D$.*

We prove Theorem 26 using the following lemma. We are able to prove a stronger version of this lemma for the special case of Averaged Two-Bin Calibration Error (ATB) in Lemma 48.

**Lemma 27** *Let $J$ be an arbitrary distribution of $(v, y) \in [0, 1] \times \{0, 1\}$, and define random variable $\widehat{v} := \mathbb{E}_J[y|v]$ as a function of $v$. Let* CAL *be an arbitrary Unnormalized Binned Squared Error (UBSE). For $T$ independent and identically distributed (i.i.d.) examples $(\widehat{v}_1, v_1, y_1), \ldots, (\widehat{v}_T, v_T, y_T)$, defining $\widehat{\boldsymbol{v}} := (\widehat{v}_1, \ldots, \widehat{v}_T)$, $\boldsymbol{v} := (v_1, \ldots, v_T)$ and $\boldsymbol{y} = (y_1, \ldots, y_T)$, we have*

$$\mathbb{E}[\textsc{Cal}(\widehat{\boldsymbol{v}}, \boldsymbol{y})] = \frac{1}{T}\,\mathbb{E}_J[(\widehat{v} - y)^2], \tag{26}$$

$$\mathbb{E}[\textsc{Cal}(\boldsymbol{v}, \boldsymbol{y})] \ge \mathbb{E}[\textsc{Cal}(\widehat{\boldsymbol{v}}, \boldsymbol{y})] + \frac{1}{T^2}\,\mathbb{E}_J[(\widehat{v} - v)^2]. \tag{27}$$

**Proof** By the definition of $\widehat{v} := \mathbb{E}_J[y|v]$, the distribution of $(\widehat{v}, y)$ is calibrated. Therefore, $\mathbb{E}[y_t|\widehat{v}_t] = \widehat{v}_t$ for every $t = 1, \ldots, T$. By Lemma 25, we have

$$\mathbb{E}[\textsc{Cal}(\widehat{\boldsymbol{v}}, \boldsymbol{y})] = \frac{1}{T^2}\,\mathbb{E}\left[\sum_{t=1}^{T} \widehat{v}_t(1 - \widehat{v}_t)\right] = \frac{1}{T}\,\mathbb{E}[\widehat{v}(1 - \widehat{v})] = \frac{1}{T}\,\mathbb{E}[(\widehat{v} - y)^2],$$

where we use the fact that $\widehat{v}_1, \ldots, \widehat{v}_T$ are independent and identically distributed (i.i.d.) random variables. This completes the proof of (26).

By Lemma 25 again, we have

$$\mathbb{E}[\textsc{Cal}(\boldsymbol{v}, \boldsymbol{y})] = \mathbb{E}[\textsc{Cal}(\boldsymbol{v}, \widehat{\boldsymbol{v}})] + \frac{1}{T^2}\,\mathbb{E}\left[\sum_{t=1}^{T} \widehat{v}_t(1 - \widehat{v}_t)\right] = \mathbb{E}[\textsc{Cal}(\boldsymbol{v}, \widehat{\boldsymbol{v}})] + \mathbb{E}[\textsc{Cal}(\widehat{\boldsymbol{v}}, \boldsymbol{y})]. \tag{28}$$

Since CAL is an Unnormalized Binned Squared Error (UBSE) (Definition 23) and there are at most $T$ non-empty bins, by the Cauchy-Schwarz inequality, we have

$$\textsc{Cal}(\boldsymbol{v}, \widehat{\boldsymbol{v}}) \ge \frac{1}{T^3}\left(\sum_{t=1}^{T}(v_t - \widehat{v}_t)\right)^2.$$

Taking expectation, we get

$$\begin{aligned}\mathbb{E}[\textsc{Cal}(\boldsymbol{v}, \widehat{\boldsymbol{v}})] &\geq \frac{1}{T^3}\left(T\,\mathbb{E}[(v-\widehat{v})^2] + T(T-1)\,\mathbb{E}[v-\widehat{v}]^2\right)\\ &\text{(because } (v_1,\widehat{v}_1),\ldots,(v_T,\widehat{v}_T) \text{ are independent and identically distributed (i.i.d.))}\\ &\geq \frac{1}{T^2}\,\mathbb{E}[(v-\widehat{v})^2].\end{aligned}$$

Plugging this into (28) proves (27). ■

We are now ready to prove Theorem 26.

**Proof** [Proof of Theorem 26] Define predictor $\widehat{r} : X \to [0,1]$ such that $\widehat{r}(x) = \mathbb{E}_D[y|r(x)]$. By Lemma 27,

$$\begin{aligned}\mathbb{E}_S[\textsc{Cal}_{D_S}(p)] &= \frac{1}{T}\,\mathbb{E}_D[(p(x)-y)^2],\\ \mathbb{E}_S[\textsc{Cal}_{D_S}(r)] &\geq \frac{1}{T}\,\mathbb{E}_D[(\widehat{r}(x)-y)^2] + \frac{1}{T^2}\,\mathbb{E}_D[(r(x)-\widehat{r}(x))^2].\end{aligned}$$

Since $\mathbb{E}[y|x] = p(x)$, we have

$$\begin{aligned}\mathbb{E}_D[(\widehat{r}(x)-y)^2] &= \mathbb{E}[(\widehat{r}(x)-p(x))^2] + \mathbb{E}[(p(x)-y)^2] + 2\,\mathbb{E}[(\widehat{r}(x)-p(x))(p(x)-y)]\\ &= \mathbb{E}[(\widehat{r}(x)-p(x))^2] + \mathbb{E}[(p(x)-y)^2].\end{aligned}$$

Therefore,

$$\begin{aligned}0 &\geq \mathbb{E}_S[\textsc{Cal}_{D_S}(r)] - \mathbb{E}_S[\textsc{Cal}_{D_S}(p)]\\ &\geq \frac{1}{T}(\mathbb{E}[(\widehat{r}(x)-y)^2] - \mathbb{E}[(p(x)-y)^2]) + \frac{1}{T^2}\,\mathbb{E}[(r(x)-\widehat{r}(x))^2]\\ &= \frac{1}{T}\,\mathbb{E}[(\widehat{r}(x)-p(x))^2] + \frac{1}{T^2}\,\mathbb{E}[(r(x)-\widehat{r}(x))^2].\end{aligned}$$

This implies that $r(x) = \widehat{r}(x) = p(x)$ almost surely. ■

## Appendix C. Calibration Errors with Two Bins

In this section, we formally define our calibration measure: the *Averaged Two-Bin Calibration Error (ATB)*. We show that ATB satisfies the following properties in the literature: completeness and soundness as a distributional calibration measure, truthfulness, continuity, consistency of the plug-in estimator, sample complexity, and computational efficiency. Our proof of the distributional completeness and soundness relies heavily on the quadratic relationship between ATB and its $\ell_1$ variant ($\ell_1$-ATB). We will show that $\ell_1$-ATB linearly approximates existing calibration measures, implying the completeness and soundness of both $\ell_1$-ATB and ATB.

**Definition 28** *For any distribution $J$ of prediction-state pairs $(v, y) \in [0,1] \times \{0,1\}$, we define the* averaged two-bin calibration error (ATB) *and its $\ell_1$ variant as follows:*

$$\begin{aligned}\mathrm{ATB}(J) &= \mathbb{E}_{q\sim\mathrm{Unif}([0,1])}\left[\left(\mathbb{E}_J\left[(v-y)\mathbb{I}\left[v<q\right]\right]\right)^2 + \left(\mathbb{E}_J\left[(v-y)\mathbb{I}\left[v\geq q\right]\right]\right)^2\right],\\ \ell_1\text{-}\mathrm{ATB}(J) &= \mathbb{E}_{q\sim\mathrm{Unif}([0,1])}\left[\left|\mathbb{E}_J\left[(v-y)\mathbb{I}\left[v<q\right]\right]\right| + \left|\mathbb{E}_J\left[(v-y)\mathbb{I}\left[v\geq q\right]\right]\right|\right].\end{aligned}$$

*Correspondingly, for any prediction sequence* $\boldsymbol{r} \in [0,1]^T$ *and ground-truth sequence* $\boldsymbol{p} \in [0,1]^T$,

$$\mathrm{ATB}(\boldsymbol{r},\boldsymbol{p}) = \mathbb{E}_{q\sim\mathrm{Unif}([0,1])}\left[\frac{1}{T^2}\left(\left(\sum_{t:r_t<q}(r_t-p_t)\right)^2+\left(\sum_{t:r_t\geq q}(r_t-p_t)\right)^2\right)\right], \tag{29}$$

$$\ell_1\text{-}\mathrm{ATB}(\boldsymbol{r},\boldsymbol{p}) = \mathbb{E}_{q\sim\mathrm{Unif}([0,1])}\left[\frac{1}{T}\left(\left|\sum_{t:r_t<q}(r_t-p_t)\right|+\left|\sum_{t:r_t\geq q}(r_t-p_t)\right|\right)\right].$$

To prepare for the proof, $\ell_1$-ATB is quadratically related to ATB by Jensen's inequality.

**Lemma 29** *For any distribution* $J$ *of prediction-state pairs* $(v,y) \in [0,1]\times\{0,1\}$,

$$\frac{1}{2}\,\ell_1\text{-}\mathrm{ATB}(J)^2 \leq \mathrm{ATB}(J) \leq \ell_1\text{-}\mathrm{ATB}(J).$$

**Proof** Fix a threshold $q$, we write $\Delta_1(q) = \mathbb{E}_J[(v-y)\mathbb{I}[v<q]]$ and $\Delta_2(q) = \mathbb{E}_J[(v-y)\mathbb{I}[v\geq q]]$. The right inequality follows from the fact that $\Delta_1, \Delta_2 \in [-1,1]$.

Using Jensen's inequality, we get the left inequality:

$$\frac{1}{2}\ell_1\text{-}\mathrm{ATB}(J)^2 = 2\left(\mathbb{E}_q\left[\frac{1}{2}|\Delta_1(q)|+\frac{1}{2}|\Delta_2(q)|\right]\right)^2 \leq 2\mathbb{E}_q\left[\frac{1}{2}\Delta_1(q)^2+\frac{1}{2}\Delta_2(q)^2\right] = \mathrm{ATB}(J).$$

■

### C.1. Completeness, Soundness, and Estimation

The distributional completeness and soundness of Averaged Two-Bin Calibration Error (ATB) follow from the quadratic approximation of the lower distance to calibration.

**Theorem 30** *Both* ATB *and* $\ell_1$-ATB *are complete and sound. Moreover, their plug-in estimators are consistent.*

Appendix C.1.1 will prove that $\ell_1$-ATB is a constant approximation of lower distance to calibration. Combined with Lemma 29, ATB is quadratically related to lower distance to calibration, which implies the completeness and soundness part of Theorem 30. The consistency of the plug-in estimators follows from the sample-complexity bound in Theorem 37.

#### C.1.1. APPROXIMATING THE DISTANCE TO CALIBRATION USING TWO BINS

In this section, we show that the $\ell_1$ variant of Averaged Two-Bin Calibration Error ($\ell_1$-ATB) is a constant-factor approximation of both smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL) (recall Proposition 9 that SMCAL and DISTCAL are constant-factor approximations to each other):

**Theorem 31** *For any distribution* $J$ *of prediction-state pairs* $(v,y) \in [0,1]\times\{0,1\}$, *we have*

$$\frac{2}{3}\,\mathrm{SMCAL}(J) \leq \ell_1\text{-}\mathrm{ATB}(J) \leq 3\,\mathrm{DISTCAL}(J).$$

Combining Theorem 31 with Proposition 9 and Lemma 29, we have the following corollary about the relationship between Averaged Two-Bin Calibration Error (ATB), its $\ell_1$ variant ($\ell_1$-ATB), smooth calibration error (SMCAL), and lower distance to calibration (DISTCAL):

**Corollary 32** *For any distribution $J$ of prediction-state pairs $(v, y) \in [0,1] \times \{0,1\}$, we have*

$$\frac{1}{3}\,\underline{\text{DISTCAL}}(J) \le \frac{2}{3}\,\text{SMCAL}(J) \le \ell_1\text{-ATB}(J) \le 3\,\underline{\text{DISTCAL}}(J) \le 6\,\text{SMCAL}(J),$$
$$\frac{1}{18}\,\underline{\text{DISTCAL}}(J)^2 \le \frac{2}{9}\,\text{SMCAL}(J)^2 \le \text{ATB}(J) \le 3\,\underline{\text{DISTCAL}}(J) \le 6\,\text{SMCAL}(J).$$

We prove the two inequalities in Theorem 31 in two separate lemmas below. We start with the easier one showing the upper bound on $\ell_1$-ATB:

**Lemma 33** *For any distribution $J$ of $(v, y) \in [0,1] \times \{0,1\}$,*

$$\ell_1\text{-ATB}(J) \le 3\,\underline{\text{DISTCAL}}(J).$$

**Proof** Let $\Pi$ be an arbitrary distribution of $(u, v, y) \in [0,1] \times [0,1] \times \{0,1\}$, where the distribution of $(v, y)$ is $J$, and the distribution of $(u, y)$ (denoted by $\widehat{J}$) is calibrated. Since $\widehat{J}$ is calibrated, we have

$$\ell_1\text{-ATB}(\widehat{J}) = 0.$$

By Theorem 36,

$$\ell_1\text{-ATB}(J) = \ell_1\text{-ATB}(J) - \ell_1\text{-ATB}(\widehat{J}) \le 3\,\mathbb{E}_\Pi\,|u - v|.$$

The lemma is proved by taking the infimum over $\Pi$. ■

Now we prove the other inequality in Theorem 31 showing the lower bound on $\ell_1$-ATB. It turns out to be convenient to first focus on the setting with $T$ fixed individuals:

**Lemma 34** *For any prediction sequence $\boldsymbol{r} \in [0,1]^T$ and any state sequence $\boldsymbol{y} \in \{0,1\}^T$, we have*

$$\text{SMCAL}(\boldsymbol{r}, \boldsymbol{y}) \le \frac{3}{2} \cdot \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}).$$

**Proof** [Proof of Lemma 34] It suffices to prove that for any 1-Lipschitz function $w : [0,1] \to [-1,1]$,

$$\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)w(r_t) \le \frac{3}{2} \cdot \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}). \tag{30}$$

Assume without loss of generality that the predictions are sorted: $r_1 \le \cdots \le r_T$. Define $w(r_0) = 0, w(r_{T+1}) = 0$. For $t = 0, \ldots, T$, define $\Delta_t := w(r_{t+1}) - w(r_t)$. We have

$$w(r_t) = \frac{1}{2}((w(r_t) - w(r_0)) - (w(r_{T+1}) - w(r_t))) = \frac{1}{2}\left(\sum_{s<t}\Delta_s - \sum_{s\ge t}\Delta_s\right)$$
$$= \frac{1}{2}\sum_{s=0}^{T}\Delta_s \mathsf{sign}(t - s),$$

where $\mathsf{sign}(u) = 1$ if $u > 0$, and $\mathsf{sign}(u) = -1$ if $u \le 0$. Therefore,

$$\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)w(r_t) = \frac{1}{2T}\sum_{s=0}^{T}\sum_{t=1}^{T}(r_t - y_t)\Delta_s \mathsf{sign}(t-s). \tag{31}$$

For $s = 1, \ldots, T-1$, by the Lipschitzness of $w$, we have $|\Delta_s| \le r_{s+1} - r_s$. Therefore,

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)\Delta_s \mathsf{sign}(t-s)\right| \\
&\le (r_{s+1} - r_s)\left|\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)\mathsf{sign}(t-s)\right| \\
&\le (r_{s+1} - r_s)\cdot\frac{1}{T}\left(\left|\sum_{t\le s}(r_t - y_t)\right| + \left|\sum_{t>s}(r_t - y_t)\right|\right) \\
&= \mathbb{E}_{q\sim\mathrm{Unif}([0,1])}\left[\mathbb{I}_{q\in[r_s, r_{s+1}]}\cdot\frac{1}{T}\left(\left|\sum_{t:r_t<q}(r_t - y_t)\right| + \left|\sum_{t:r_t\ge q}(r_t - y_t)\right|\right)\right].
\end{aligned}$$

Summing up over $s = 1, \ldots T-1$, we have

$$\sum_{s=1}^{T-1}\left|\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)\Delta_s \mathsf{sign}(t-s)\right| \le \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}). \tag{32}$$

Moreover, since $w(r_1), w(r_T) \in [-1, 1]$, we have $|\Delta_0|, |\Delta_T| \le 1$. Therefore,

$$\left|\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)\Delta_0 \mathsf{sign}(t-0)\right| = |\Delta_0|\cdot\left|\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)\right| \le \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}), \tag{33}$$

$$\left|\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)\Delta_T \mathsf{sign}(t-T)\right| = |\Delta_T|\cdot\left|\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)\right| \le \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}). \tag{34}$$

Adding up the three inequalities (32) (33) (34) above, we get

$$\sum_{s=0}^{T}\left|\frac{1}{T}\sum_{t=1}^{T}(r_t - y_t)\Delta_s \mathsf{sign}(t-s)\right| \le 3\,\ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}).$$

Combining this with (31) using the triangle inequality, we get (30), as desired. ■

**Proof** [Proof of Theorem 31] The upper bound on $\ell_1$-ATB has been proved in Lemma 33. It remains to establish the lower bound on $\ell_1$-ATB in terms of smooth calibration error (SMCAL):

$$\textsc{smCal}(J) \le \frac{3}{2}\cdot\ell_1\text{-ATB}(J). \tag{35}$$

Consider a sample $S$ of $T$ independent and identically distributed (i.i.d.) points $(v_1, y_1), \ldots, (v_T, y_T)$ from $J$. Defining $\boldsymbol{r} := (v_1, \ldots, v_T)$ and $\boldsymbol{y} := (y_1, \ldots, y_T)$, we have

$$\begin{aligned}\textsc{smCal}(J_S) &= \textsc{smCal}(J_{\boldsymbol{r},\boldsymbol{y}}) = \textsc{smCal}(\boldsymbol{r}, \boldsymbol{y}),\\ \ell_1\text{-ATB}(J_S) &= \ell_1\text{-ATB}(J_{\boldsymbol{r},\boldsymbol{y}}) = \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}),\end{aligned}$$

where we use the $J_{\boldsymbol{r},\boldsymbol{y}}$ notation from Definition 18. By Lemma 34,

$$\textsc{smCal}(J_S) = \textsc{smCal}(\boldsymbol{r}, \boldsymbol{y}) \leq \frac{3}{2} \cdot \ell_1\text{-ATB}(\boldsymbol{r}, \boldsymbol{y}) = \frac{3}{2} \cdot \ell_1\text{-ATB}(J_S). \tag{36}$$

Taking $T \to \infty$, by Theorem 37 and Proposition 38, we know that $\textsc{smCal}(J_S)$ converges in probability to $\textsc{smCal}(J)$, and that $\ell_1\text{-ATB}(J_S)$ converges in probability to $\ell_1\text{-ATB}(J)$. Therefore, our goal (35) follows from (36). ■

### C.2. (Strict) Truthfulness

From its definition (29), Averaged Two-Bin Calibration Error (ATB) is clearly a special case of an Unnormalized Binned Squared Error (UBSE) (Definition 23), so its truthfulness follows immediately from Theorem 24 and Theorem 26.

**Theorem 35 (Truthfulness)** *The calibration measure* ATB *is interim truthful (Definition 19). Moreover,* ATB *inherits the error decomposition:*

$$\mathbb{E}_{\boldsymbol{y}\sim\boldsymbol{p}}[\text{ATB}(\boldsymbol{r}, \boldsymbol{y})] = \text{ATB}(\boldsymbol{r}, \boldsymbol{p}) + \frac{1}{T^2}\sum_{t=1}^{T} p_t(1 - p_t).$$

*Moreover,* ATB *is strictly ex-ante truthful.*

### C.3. Continuity

The following theorem establishes the continuity of ATB and $\ell_1$-ATB with a general formalization. Both errors change continuously as the predictions change.

**Theorem 36 (Continuity)** *Let* $\Pi$ *be a joint distribution of* $(v_1, v_2, y) \in [0,1] \times [0,1] \times \{0,1\}$. *Let* $J_1$ *denote the distribution of* $(v_1, y)$, *and let* $J_2$ *denote the distribution of* $(v_2, y)$. *We have*

$$|\ell_1\text{-ATB}(J_1) - \ell_1\text{-ATB}(J_2)| \leq 3\ \mathbb{E}_{\Pi}\,|v_1 - v_2|. \tag{37}$$

$$|\text{ATB}(J_1) - \text{ATB}(J_2)| \leq 6\ \mathbb{E}_{\Pi}\,|v_1 - v_2|. \tag{38}$$

**Proof** By Definition 28, we have

$$\begin{aligned}\ell_1\text{-ATB}(J_1) &= \mathbb{E}_{q\sim\text{Unif}([0,1])}\left[\left|\mathbb{E}_{\Pi}\left[(v_1 - y)\mathbb{I}\left[v_1 < q\right]\right]\right| + \left|\mathbb{E}_{\Pi}\left[(v_1 - y)\mathbb{I}\left[v_1 \geq q\right]\right]\right|\right],\\ \ell_1\text{-ATB}(J_2) &= \mathbb{E}_{q\sim\text{Unif}([0,1])}\left[\left|\mathbb{E}_{\Pi}\left[(v_2 - y)\mathbb{I}\left[v_2 < q\right]\right]\right| + \left|\mathbb{E}_{\Pi}\left[(v_2 - y)\mathbb{I}\left[v_2 \geq q\right]\right]\right|\right].\end{aligned}$$

We define an intermediate quantity

$$\kappa := \mathbb{E}_{q\sim\text{Unif}([0,1])}\left[\left|\mathbb{E}_{\Pi}\left[(v_2 - y)\mathbb{I}\left[v_1 < q\right]\right]\right| + \left|\mathbb{E}_{\Pi}\left[(v_2 - y)\mathbb{I}\left[v_1 \geq q\right]\right]\right|\right].$$

By the triangle inequality,

$$\begin{aligned}|\ell_1\text{-ATB}(J_1) - \kappa| &\le \mathbb{E}_{q\sim\text{Unif}([0,1])}\left[\mathbb{E}_{\Pi}\left[\left|v_1 - v_2\right|\mathbb{I}\left[v_1 < q\right]\right] + \mathbb{E}_{\Pi}\left[\left|v_1 - v_2\right|\mathbb{I}\left[v_1 \ge q\right]\right]\right] \\ &= \mathbb{E}_{\Pi}\left|v_1 - v_2\right|. \end{aligned} \tag{39}$$

Similarly, noting that $|v_2 - y| \le 1$, we have

$$\begin{aligned}|\ell_1\text{-ATB}(J_2) - \kappa| &\le \mathbb{E}_{q\sim\text{Unif}([0,1])}\left[2\ \mathbb{E}_{\Pi}\left|\mathbb{I}\left[v_1 < q\right] - \mathbb{I}\left[v_2 < q\right]\right|\right] \\ &= 2\,\mathbb{E}_{\Pi}\left[\mathbb{E}_{q\sim\text{Unif}([0,1])}\left|\mathbb{I}\left[v_1 < q\right] - \mathbb{I}\left[v_2 < q\right]\right|\right] \\ &= 2\,\mathbb{E}_{\Pi}\left|v_1 - v_2\right|. \end{aligned} \tag{40}$$

Summing up (39) and (40) proves (37). A similar strategy proves (38), using one extra observation: the function $u^2$ is 2-Lipshitz for $u \in [-1, 1]$. We omit the details. ■

### C.4. Sample Complexity

Both Averaged Two-Bin Calibration Error (ATB) and its $\ell_1$ variant ($\ell_1$-ATB) can be estimated within error $\varepsilon$ using $O(1/\varepsilon^2)$ independent and identically distributed (i.i.d.) examples:

**Theorem 37 (Sample complexity)** *Let $J$ be any distribution of prediction-state pairs $(v, y) \in [0, 1] \times \{0, 1\}$, and let $S$ be a sample of $T$ i.i.d. points $(v_1, y_1), \ldots, (v_T, y_T)$ from $J$. For $\varepsilon, \delta \in (0, 1/3)$, assume $T > C\varepsilon^{-2}\log(1/\delta)$ for a sufficiently large absolute constant $C > 0$. With probability at least $1 - \delta$ (over the randomness in the sample $S$),*

$$\begin{aligned}|\ell_1\text{-ATB}(J_S) - \ell_1\text{-ATB}(J)| &\le \varepsilon, \\ |\text{ATB}(J_S) - \text{ATB}(J)| &\le \varepsilon.\end{aligned}$$

**Proof** It suffices to show that with probability at least $1 - \delta$, for every $q \in [0, 1]$,

$$\begin{aligned}&\left|\mathbb{E}_{J_S}\left[(v - y)\mathbb{I}\left[v < q\right]\right] - \mathbb{E}_{J}\left[(v - y)\mathbb{I}\left[v < q\right]\right]\right| \le \varepsilon/4, \quad \text{and} \\ &\left|\mathbb{E}_{J_S}\left[(v - y)\mathbb{I}\left[v \ge q\right]\right] - \mathbb{E}_{J}\left[(v - y)\mathbb{I}\left[v \ge q\right]\right]\right| \le \varepsilon/4.\end{aligned}$$

By Proposition 50, it suffices to prove the following Rademacher complexity bounds for the function families $F = \{f_q\}_{q\in[0,1]}$ and $G = \{g_q\}_{q\in[0,1]}$ where $f_q(v, y) = (v - y)\mathbb{I}\left[v < q\right]$ and $g_q(v, y) = (v - y)\mathbb{I}\left[v \ge q\right]$: for every $(v_1, y_1), \ldots, (v_T, y_T) \in [0, 1] \times \{0, 1\}$,

$$\mathcal{R}(F; (v_1, y_1), \ldots, (v_T, y_T)) \le O\left(\sqrt{\frac{1}{T}}\right), \quad \text{and} \tag{41}$$

$$\mathcal{R}(G; (v_1, y_1), \ldots, (v_T, y_T)) \le O\left(\sqrt{\frac{1}{T}}\right). \tag{42}$$

Now consider the family $H = \{h_q\}_{q\in[0,1]}$ where $h_q(v, y) = \mathbb{I}\left[v < q\right]$. Clearly, $H$ has VC dimension at most 1. By Proposition 53, we have

$$\mathcal{R}(H; (v_1, y_1), \ldots, (v_T, y_T)) \le O\left(\sqrt{\frac{1}{T}}\right). \tag{43}$$

Observe that $f_q(v_i, y_i) = \eta_i(h_q(v_i, y_i))$ for every $i = 1, \dots, T$ and $q \in [0, 1]$, where $\eta_i$ is the univariate function $\eta_i(u) = (v_i - y_i)u$ for $u \in \mathbb{R}$. Since $|v_i - y_i| \le 1$, the function $\eta_i$ is 1-Lipschitz. Therefore, by Proposition 51, inequality (41) follows from (43). Inequality (42) can be proved similarly. ■

We remark that an analogous sample complexity bound for smooth calibration error (SMCAL) has been shown by Błasiok et al. (2023) using a similar analysis:

**Proposition 38 (Błasiok et al. (2023))** *Let $J$ be any distribution of prediction-state pairs $(v, y) \in [0, 1] \times \{0, 1\}$, and let $S$ be a sample of $T$ independent and identically distributed (i.i.d.) points $(v_1, y_1), \dots, (v_T, y_T)$ from $J$. For $\varepsilon, \delta \in (0, 1/3)$, assume $T > C\varepsilon^{-2} \log(1/\delta)$ for a sufficiently large absolute constant $C > 0$. With probability at least $1 - \delta$ (over the randomness in the sample $S$),*

$$|\text{SMCAL}(J_S) - \text{SMCAL}(J)| \le \varepsilon.$$

### C.5. Computational Efficiency

As we show in the following theorem, Averaged Two-Bin Calibration Error (ATB) can be computed and approximated efficiently.

**Theorem 39** *Given $\boldsymbol{r}, \boldsymbol{p} \in [0, 1]^T$, we can compute $\text{ATB}(\boldsymbol{r}, \boldsymbol{p})$ in time $O(T \log T)$. We can also approximate $\text{ATB}(\boldsymbol{r}, \boldsymbol{p})$ up to arbitrary additive error $\varepsilon > 0$ in time $O(T + 1/\varepsilon)$.*

The algorithm we use to prove Theorem 39 is extremely easy to describe and implement. Define

$$\Delta_1(q) = \frac{1}{T}\left|\sum_{t:r_t<q}(r_t - p_t)\right| \quad \text{and} \quad \Delta_2(q) = \frac{1}{T}\left|\sum_{t:r_t\ge q}(r_t - p_t)\right|.$$

The following algorithm computes ATB:

- $O(T \log T)$ time: sort predictions in increasing order such that $r_1 \le r_2 \le \cdots \le r_T$. Define $r_0 = 0$ and $r_{T+1} = 1$.
- $O(T)$ time: for $q = r_1, \dots, r_{T+1}$, calculate $\Delta_1(q)$ by scanning predictions in increasing order. Similarly, calculate $\Delta_2(q)$ by scanning predictions in decreasing order.
- $O(T)$ time: Calculate the expectation over threshold $q$: for $t = 1, \dots, T + 1$, sum up $\Delta_1(r_t)^2 + \Delta_2(r_t)^2$ with weight $|r_t - r_{t-1}|$.

If we allow additive error $\varepsilon \in (0, 1)$, by Theorem 36, we can round the predictions $r_1, \dots, r_T$ to multiples of $\varepsilon/6$ and then compute ATB exactly. This makes all predictions $r_1, \dots, r_T$ lie in a finite set $\{0, \varepsilon/6, 2\varepsilon/6, \dots\} \cap [0, 1]$ of size $O(1/\varepsilon)$, so the sorting step can be implemented in time $O(T + 1/\varepsilon)$ by bucket sort.

A similar algorithm computes $\ell_1$-$\text{ATB}(\boldsymbol{r}, \boldsymbol{p})$ in $O(T \log T)$ time, or approximates $\ell_1$-$\text{ATB}(\boldsymbol{r}, \boldsymbol{p})$ up to error $\varepsilon$ in $O(T + 1/\varepsilon)$ time. We note that currently known algorithms for computing smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL) are much more complicated, with the best known running time being $O(T \log^2 T)$ and $O(T^2 \log T)$, respectively, even when $O(1/\sqrt{T})$ additive error is allowed (Hu et al., 2024).

## Appendix D. Linear-Time Calibration Tester

In this section, we show that Averaged Two-Bin Calibration Error (ATB) and its $\ell_1$ variant ($\ell_1$-ATB) are both optimally valid (Definition 22) for smooth calibration error (SMCAL) and lower distance to calibration (DISTCAL). It is fairly straightforward to show that $\ell_1$-ATB is $O(1/\sqrt{T})$-valid using its constant approximation to SMCAL (Theorem 31) and its sample complexity bound (Theorem 37). In Theorem 40 below, we show that ATB is $O(1/\sqrt{T})$-valid as well, and that this is optimal up to constant.

These results imply faster algorithms for solving the calibration testing problem studied by Hu et al. (2024), which requires distinguishing, with large constant success probability, whether a distribution $J$ is perfectly calibrated or has $\text{DISTCAL}(J) > \varepsilon$ given independent and identically distributed (i.i.d.) data points drawn from $J$. This can be solved by computing ATB or $\ell_1$-ATB on $T = O(1/\varepsilon^2)$ data points and compare the result with the threshold $\beta_T$ in the definition of validity (Definition 22). By Theorem 39, the running time we need is $O(T \log T)$, which already improves the $O(T \log^2 T)$ time bound of Hu et al. (2024). Moreover, our Lemmas 41 and 42 show that it suffices to approximate ATB up to additive error $1/(2T)$, which can be achieved in time $O(T)$ by Theorem 39, giving the first linear-time algorithm for calibration testing.

**Theorem 40** *The calibration measure Averaged Two-Bin Calibration Error (*ATB*) is $O(\frac{1}{\sqrt{T}})$-valid w.r.t. the reference calibration error lower distance to calibration (*DISTCAL*). That is,* ATB *is $\{\gamma_T\}$-valid for some sequence $\gamma_1, \gamma_2, \ldots$ with $\gamma_T = O(1/\sqrt{T})$. Moreover, this is optimal up to constant factors: if there exists a $\{\gamma_T\}$-valid calibration error w.r.t.* DISTCAL*, then $\gamma_T = \Omega(1/\sqrt{T})$.*

Theorem 40 is an immediate corollary of the following Lemmas 41, 42, and 43.

**Lemma 41** *Let $J$ be an arbitrary distribution of prediction-state pairs $(v, y) \in [0,1] \times \{0,1\}$ and assume that $J$ is calibrated. For any $T \in \mathbb{Z}_{>0}$, consider a sample $S$ of $T$ independent and identically distributed (i.i.d.) points $(v_1, y_1), \ldots, (v_T, y_T) \in [0,1] \times \{0,1\}$ from $J$, and let $J_S$ be the uniform distribution over $S$. We have*

$$\Pr_{S \sim J^T}[\text{ATB}(J_S) \leq 1/T] \geq 3/4.$$

**Proof** Define $\boldsymbol{r} = (v_1, \ldots, v_T)$ and $\boldsymbol{y} = (y_1, \ldots, y_T)$. It is clear that the distribution $J_{\boldsymbol{r},\boldsymbol{y}}$ (see Definition 18) is equal to the distribution $J_S$. Therefore,

$$\text{ATB}(J_S) = \text{ATB}(\boldsymbol{r}, \boldsymbol{y}).$$

Since $J$ is calibrated, we have $\mathbb{E}_J[y|v = v_t] = v_t$ for every $t = 1, \ldots, T$. Conditioned on $\boldsymbol{r} = (v_1, \ldots, v_T)$, each $y_t$ is independently distributed as the Bernoulli distribution with mean $v_t$. Thus, we have $\boldsymbol{y} \sim \boldsymbol{r}$ as in Definition 19. Therefore,

$$\Pr_S[\text{ATB}(J_S) \leq 1/T \mid v_1, \ldots, v_T] = \Pr_{\boldsymbol{y} \sim \boldsymbol{r}}[\text{ATB}(\boldsymbol{r}, \boldsymbol{y}) \leq 1/T]. \tag{44}$$

By Lemma 25,

$$\mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{r}}[\text{ATB}(\boldsymbol{r}, \boldsymbol{y})] = \text{ATB}(\boldsymbol{r}, \boldsymbol{r}) + \frac{1}{T^2} \sum_{t=1}^{T} p_t(1 - p_t) = \frac{1}{T^2} \sum_{t=1}^{T} p_t(1 - p_t) \leq \frac{1}{4T}.$$

By Markov's inequality,

$$\Pr_{\boldsymbol{y}\sim\boldsymbol{r}}[\mathrm{ATB}(\boldsymbol{r},\boldsymbol{y}) \le 1/T] \ge 3/4.$$

Plugging this into (44) and taking the expectation over $v_1, \ldots, v_T$ completes the proof. ■

**Lemma 42** *There exists an absolute constant $C > 0$ such that the following holds. For any $T \in \mathbb{Z}_{>0}$ and any distribution $J$ of $(v, y) \in [0,1] \times \{0,1\}$ with lower distance to calibration ($\underline{\mathrm{DISTCAL}}(J)$) at least $C/\sqrt{T}$, let $S$ be a sample of $T$ independent and identically distributed (i.i.d.) points from $J$. Then*

$$\Pr_{S\sim J^T}[\mathrm{ATB}(J_S) \le 2/T] \le 1/4.$$

**Proof** By Theorem 37, there exists an absolute constant $C' > 0$ such that with probability at least $3/4$ over $S \sim J^T$,

$$|\ell_1\text{-ATB}(J_S) - \ell_1\text{-ATB}(J)| \le C'/\sqrt{T}. \tag{45}$$

It remains to show that whenever (45) holds, we have

$$\mathrm{ATB}(J_S) > 2/T.$$

By Corollary 32 and our assumption that $\underline{\mathrm{DISTCAL}}(J) > C/\sqrt{T}$, we have $\ell_1\text{-ATB}(J) \ge (C/3)/\sqrt{T}$. Therefore, whenever (45) holds, we have

$$\ell_1\text{-ATB}(J_S) \ge (C/3 - C')/\sqrt{T}.$$

Assuming $C/3 - C' > 0$ which is guaranteed by a sufficiently large $C$, by Lemma 29, we have

$$\mathrm{ATB}(J_S) \ge (1/2)(C/3 - C')^2/T.$$

The proof is completed by choosing $C$ large enough so that $(1/2)(C/3 - C')^2 > 2$. ■

**Lemma 43** *Let $\{\gamma_T\}_{T=1,2,\ldots}$ be a sequence of nonnegative real numbers such that there exists a $\{\gamma_T\}$-valid calibration error* CAL *w.r.t.* $\underline{\mathrm{DISTCAL}}$. *Then $\gamma_T = \Omega(1/\sqrt{T})$.*

**Proof** Let us focus on the choices of $T$ such that $\gamma_T < 1/2$. We define $J_1$ to be the uniform distribution over $\{(1/2, 0), (1/2, 1)\} \subseteq [0,1] \times \{0,1\}$. We define $J_2$ to be the distribution with probability mass $1/2 - \gamma_T$ on $(1/2, 0)$, and the remaining probability mass $1/2 + \gamma_T$ on $(1/2, 1)$.

Clearly, $J_1$ is calibrated. We claim that $\underline{\mathrm{DISTCAL}}(J_2) \ge \gamma_T$. Indeed, consider any coupling distribution $\Pi$ of $(u, v, y) \in [0,1] \times [0,1] \times \{0,1\}$, where $(v, y)$ is distributed as $J_2$, and the distribution of $(u, y)$ is calibrated. By calibration, $\mathbb{E}[u] = \mathbb{E}[y] = 1/2 + \gamma_T$. Therefore, $\mathbb{E}\,|u - v| \ge \mathbb{E}[u] - \mathbb{E}[v] = \gamma_T$, implying that $\underline{\mathrm{DISTCAL}}(J_2) \ge \gamma_T$.

Let $\beta_T \in \mathbb{R}$ be the threshold satisfying the requirement of validity (Definition 22). Define

$$\delta_T := \mathsf{accP}^{\mathrm{CAL}}(J_1; T, \beta_T) - \mathsf{accP}^{\mathrm{CAL}}(J_2; T, \beta_T).$$

Note that the two acceptance probabilities above are w.r.t. the randomness in the samples $S_1 \sim J_1^T$ and $S_2 \sim J_2^T$, respectively, where $J_1^T$ (resp. $J_2^T$) is the joint distribution of $T$ i.i.d. points from $J_1$

(resp. $J_2$). A standard argument (e.g. via Pinsker's inequality) shows that the total variation distance between $J_1^T$ and $J_2^T$ is $O(\gamma_T\sqrt{T})$. Therefore,

$$\delta_T \leq O(\gamma_T\sqrt{T}).$$

Validity requires $\liminf_{T\to\infty} \delta_T > 0$. Therefore,

$$\liminf_{T\to\infty} \gamma_T\sqrt{T} > 0.$$

This implies $\gamma_T = \Omega(1/\sqrt{T})$. ■

## Appendix E. Non-Truthfulness of Known Calibration Measures

In this section, we prove Theorem 45 showing that condition (1) (restated below as condition (46)) holds for a broad family of calibration measures: $\ell_\alpha$ Expected Calibration Error ($\ell_\alpha$-ECE), $\ell_\alpha$ binned Expected Calibration Error ($\ell_\alpha$-BINECE), smooth calibration error (SMCAL), and $\ell_\alpha$ lower distance to calibration ($\ell_\alpha$-$\underline{\text{DISTCAL}}$) (Definition 44), where $\alpha \geq 1$ is arbitrary. Condition (1) states that, for any realization of the states, reporting the average over predictions is weakly better than reporting truthfully for known calibration measures. By Remark 46, this proves that these calibration measures are not truthful.

**Definition 44 ($\ell_\alpha$-Distance to Calibration)** *Let $J$ be a distribution of $(v, y) \in [0, 1] \times \{0, 1\}$. We define its $\ell_\alpha$-distance to calibration (denoted by $\ell_\alpha$-$\underline{\text{DISTCAL}}(J)$) similarly to the definition of $\underline{\text{DISTCAL}}(J)$ in Definition 7. The only difference is that we change the $\ell_1$ distance $|u-v|$ to $|u-v|^\alpha$:*

$$\ell_\alpha\text{-}\underline{\text{DISTCAL}}(J) := \inf_{\Pi} \mathbb{E}_\Pi[|u - v|^\alpha],$$

*where the infimum is over joint distributions $\Pi$ of $(u, v, y) \in [0, 1] \times [0, 1] \times \{0, 1\}$, where $(v, y)$ is distributed according to $J$, and the distribution of $(u, y)$ is calibrated.*

**Theorem 45** *Let* CAL *be a calibration measure from $\{\ell_\alpha$-ECE, $\ell_\alpha$-BINECE, SMCAL, $\ell_\alpha$-$\underline{\text{DISTCAL}}\}$, where $\alpha \geq 1$ is arbitrary. For every $\boldsymbol{r} = (r_1, \ldots, r_T) \in [0, 1]^T$ and every $\boldsymbol{y} \in \{0, 1\}^T$, it holds that*

$$\text{CAL}(\bar{\boldsymbol{r}}, \boldsymbol{y}) \leq \text{CAL}(\boldsymbol{r}, \boldsymbol{y}), \tag{46}$$

*where $\bar{\boldsymbol{r}} := (\bar{r}, \ldots, \bar{r}) \in [0, 1]^T$ for $\bar{r} := \frac{1}{T}\sum_{t=1}^T r_t$.*

**Remark 46** *It is very easy to find (many) examples of $\boldsymbol{y}$ where the inequality (46) becomes strict, in which case reporting $\bar{\boldsymbol{r}}$ instead of $\boldsymbol{r}$ is strictly better (i.e. $\bar{\boldsymbol{r}}$ dominates $\boldsymbol{r}$). In particular, we can find many examples of $\boldsymbol{y}$ where $\text{CAL}(\boldsymbol{r}, \boldsymbol{y}) > 0$ and $\text{CAL}(\bar{\boldsymbol{r}}, \boldsymbol{y}) = 0$. This is because for all of the calibration measures* CAL *mentioned above (including the continuous ones), we always have $\text{CAL}(\boldsymbol{r}, \boldsymbol{y}) > 0$ as long as $r_1, \ldots, r_T$ are distinct values[5] in, say, $[1/3, 2/3]$, and we have $\text{CAL}(\bar{\boldsymbol{r}}, \boldsymbol{y}) = 0$ as long as the average outcome $\frac{1}{T}\sum_{t=1}^T y_t$ is equal to $\bar{r}$.*

5. For binned calibration error, we need $\max_t r_t - \min_t r_t$ to be sufficiently large so that $r_1, \ldots, r_T$ do not fall in the same bin.

Theorem 45 is a corollary of the following theorem.

**Theorem 47** *Let* CAL *be a calibration measure from* $\{\ell_\alpha\text{-ECE}, \ell_\alpha\text{-BINECE}, \text{SMCAL}, \ell_\alpha\text{-}\underline{\text{DISTCAL}}\}$*, i.e., from* $\ell_\alpha$ *Expected Calibration Error (ECE),* $\ell_\alpha$ *binned Expected Calibration Error (Binned ECE), smooth calibration error (*SMCAL*), and* $\ell_\alpha$ *lower distance to calibration (*$\ell_\alpha$-$\underline{\text{DISTCAL}}$*), where* $\alpha \geq 1$ *is arbitrary. Let* $J$ *be an arbitrary distribution of* $(v, y) \in [0,1] \times \{0,1\}$*. We have*

$$\text{CAL}(\bar{J}) \leq \text{CAL}(J). \tag{47}$$

*Here* $\bar{J}$ *is the distribution of* $(\bar{v}, y) \in [0,1] \times \{0,1\}$*, where we draw* $(v, y) \sim J$ *and replace* $v$ *with the deterministic quantity* $\bar{v} := \mathbb{E}_J[v]$*.*

**Proof** We prove the theorem separately for each choice of CAL. Similarly to the definition of $\bar{v}$, we define $\bar{y} := \mathbb{E}_J[y]$.

When $\text{CAL} = \ell_\alpha$-ECE, defining $\widehat{v} := \mathbb{E}[y|v]$, by Jensen's Inequality we have

$$\text{ECE}(J) = \mathbb{E}[|v - \widehat{v}|^\alpha] \geq |\,\mathbb{E}[v - \widehat{v}]|^\alpha = |\bar{v} - \bar{y}|^\alpha, \tag{48}$$

where we used the fact that $\mathbb{E}[\widehat{v}] = \mathbb{E}[y] = \bar{y}$. Also,

$$\text{ECE}(\bar{J}) = \mathbb{E}[|\bar{v} - \mathbb{E}[y|\bar{v}]|^\alpha] = |\bar{v} - \bar{y}|^\alpha, \tag{49}$$

where we used the fact that $\bar{v}$ is a deterministic quantity, so $\mathbb{E}[y|\bar{v}] = \mathbb{E}[y] = \bar{y}$. Combining (48) and (49) proves (47).

When $\text{CAL} = \ell_\alpha$-BINECE, we can prove (47) as follows. Let $\mathcal{I} = \{I_i\}_{i\in[k]}$ be the partition of the prediction space $[0,1]$ in the definition of $\ell_\alpha$-BINECE (Definition 15). By Jensen's Inequality,

$$\begin{aligned}\ell_\alpha\text{-BINECE}(J) = \sum_{i\in[k]} \Pr_J[v \in I_i] \cdot \big|\,\mathbb{E}_J[v - y|v \in I_i]\big|^\alpha &\geq \left|\sum_{i\in[k]} \Pr_J[v \in I_i] \cdot \mathbb{E}_J[v - y|v \in I_i]\right|^\alpha \\ &= |\,\mathbb{E}_J[v - y]|^\alpha \\ &= |\bar{v} - \bar{y}|^\alpha. \end{aligned} \tag{50}$$

Also, since $\bar{v}$ is deterministic, we have

$$\ell_\alpha\text{-BINECE}(\bar{J}) = \sum_{i\in[k]} \Pr[\bar{v} \in I_i] \cdot \big|\,\mathbb{E}[\bar{v} - y|\bar{v} \in I_i]\big|^\alpha = |\,\mathbb{E}[\bar{v} - y]|^\alpha = |\bar{v} - \bar{y}|^\alpha. \tag{51}$$

Combining (50) and (51) proves (47).

When $\text{CAL} = \text{SMCAL}$, (47) follows from the following calculation:

$$\begin{aligned}\text{SMCAL}(\bar{J}) &= \sup_{w\in W_1} \mathbb{E}[(\bar{v} - y)w(\bar{v})] = \sup_{w\in W_1} (\bar{v} - \bar{y})w(\bar{v}) = |\bar{v} - \bar{y}|, \\ \text{SMCAL}(J) &= \sup_{w\in W_1} \mathbb{E}[(v - y)w(v)] \geq \sup_{\sigma\in\{\pm 1\}} \mathbb{E}[(v - y)\sigma] = \sup_{\sigma\in\{\pm 1\}} (\bar{v} - \bar{y})\sigma = |\bar{v} - \bar{y}|.\end{aligned}$$

When $\text{CAL} = \ell_\alpha$-$\underline{\text{DISTCAL}}$, (47) holds because of the following argument. Consider any joint distribution $\Pi$ of $(u, v, y) \in [0,1] \times [0,1] \times \{0,1\}$, where the marginal distribution of $(v, y)$ is $J$, and the marginal distribution of $(u, y)$ is calibrated. By Jensen's Inequality,

$$\mathbb{E}_\Pi[|u - v|^\alpha] \geq |\,\mathbb{E}_\Pi[u - v]|^\alpha = |\,\mathbb{E}_\Pi[u] - \mathbb{E}_\Pi[v]|^\alpha = |\,\mathbb{E}_\Pi[y] - \mathbb{E}_\Pi[v]]|^\alpha = |\bar{y} - \bar{v}|^\alpha.$$

| Prob. | States | SMCAL | | DISTCAL | | $\ell_2$-DISTCAL | | ATB (ours) | |
|---|---|---|---|---|---|---|---|---|---|
| | | avg | truth | avg | truth | avg | truth | avg | truth |
| $\frac{3}{16}$ | $(0,0)$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.25 | 0.3125 | 0.25 | 0.203125 |
| $\frac{3}{16}$ | $(1,1)$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.25 | 0.3125 | 0.25 | 0.203125 |
| $\frac{9}{16}$ | $(0,1)$ | 0 | 0.0625 | 0 | $> 0$ | 0 | $> 0$ | 0 | 0.015625 |
| $\frac{1}{16}$ | $(1,0)$ | 0 | 0.1875 | 0 | $> 0$ | 0 | $> 0$ | 0 | 0.140625 |
| Expected Error | | 0.1875 | 0.234375 | 0.1875 | $> 0.1875$ | 0.09375 | $> 0.11$ | **0.09375** | **0.09375** |

Table 1: The calibration errors of predictors with two data points. The ground truth probabilities of the two points are $25\%$ and $75\%$, respectively. In the table, `avg` stands for the uninformative predictor that always outputs $50\%$ and `truth` stands for the truthful predictor that outputs $25\%$ and $75\%$. We calculate the error of the predictors given each realization of the state and the total expected error. For non-truthful error metrics, the expected error of a truthful predictor is strictly higher than the expected error of an uninformative predictor. For ATB, the expected errors are the same.

We used the fact that $\mathbb{E}_\Pi[u] = \mathbb{E}_\Pi[y]$, which holds because the distribution of $(u, y)$ is calibrated. Taking the infimum over $\Pi$, we get

$$\underline{\text{DISTCAL}}(J) \geq |\bar{y} - \bar{v}|^\alpha. \tag{52}$$

Moreover, since always predicting $\bar{y}$ yields a calibrated predictor, we can choose $\Pi$ to be the joint distribution of $(\bar{y}, \bar{v}, y)$ and get

$$\underline{\text{DISTCAL}}(\bar{J}) \leq \mathbb{E}_\Pi[|\bar{y} - \bar{v}|^\alpha] = |\bar{v} - \bar{y}|^\alpha. \tag{53}$$

Combining (52) and (53) proves (47). ■

Table 1 provides an example illustrating the non-truthfulness of known calibration measures and the truthfulness of our Averaged Two-Bin Calibration Error (ATB). The table compares two strategies: predicting the overall average and predicting truthfully.

## Appendix F. Averaged Two-Bin Calibration Error (ATB) and Brier Loss

For Averaged Two-Bin Calibration Error (ATB) as a special case of an Unnormalized Binned Squared Error (UBSE), we have the following stronger result than Lemma 27:

**Lemma 48 (Averaged Two-Bin Calibration Error (ATB) and Brier loss)** *Let $J$ be an arbitrary distribution of $(v, y) \in [0,1] \times \{0,1\}$. For $(v_1, y_1), \ldots, (v_T, y_T)$ drawn as independent and identically distributed (i.i.d.) samples from $J$, defining $\boldsymbol{v} := (v_1, \ldots, v_T), \boldsymbol{y} = (y_1, \ldots, y_T)$, we have*

$$\mathbb{E}[\text{ATB}(\boldsymbol{v}, \boldsymbol{y})] \geq \frac{1}{T}\, \mathbb{E}_{(v,y)\sim J}[(v - y)^2]. \tag{54}$$

*The inequality becomes an equality if $J$ is perfectly calibrated.*

**Proof** For every fixed threshold $q \in [0,1]$, we have

$$\mathbb{E}\left[\left(\sum_{t:v_t<q}(v_t-y_t)\right)^2\right] = \mathbb{E}\left[\left(\sum_{t=1}^{T}(v_t-y_t)\mathbb{I}\left[v_t<q\right]\right)^2\right]$$
$$= T\,\mathbb{E}_J[((v-y)\mathbb{I}\left[v<q\right])^2] + T(T-1)\,\mathbb{E}_J[(v-y)\mathbb{I}\left[v<q\right]]^2$$

(because $(v_1,y_1),\ldots,(v_T,y_T)$ are drawn as independent and identically distributed (i.i.d.) samples from $J$)

$$\geq T\,\mathbb{E}_J[((v-y)\mathbb{I}\left[v<q\right])^2]$$
$$= T\,\mathbb{E}_J[(v-y)^2\mathbb{I}\left[v<q\right]].$$

Similarly,

$$\mathbb{E}\left[\left(\sum_{t:v_t\geq q}(v_t-y_t)\right)^2\right] \geq T\,\mathbb{E}_J[(v-y)^2\mathbb{I}\left[v\geq q\right]].$$

Summing up the two inequalities above, for every $q \in [0,1]$ we have

$$\frac{1}{T^2}\,\mathbb{E}\left[\left(\sum_{t:v_t<q}(v_t-y_t)\right)^2 + \left(\sum_{t:v_t\geq q}(v_t-y_t)\right)^2\right] \geq \frac{1}{T}\,\mathbb{E}_J[(v-y)^2].$$

Taking expectation over $q \sim \mathrm{Unif}([0,1])$ proves (54). When $J$ is perfectly calibrated, all inequalities in this proof become equalities, so (54) also holds as an equality. ■

## Appendix G. Standard Uniform Convergence Bounds

We include some standard notions and results on concentration inequalities and sample complexity bounds for uniform convergence. They are used when we prove the sample complexity bounds for estimating Averaged Two-Bin Calibration Error (ATB) and its $\ell_1$ variant ($\ell_1$-ATB) in Theorem 37.

We start with the definition of the Rademacher complexity.

**Definition 49 (Rademacher complexity)** *Let $F$ be a family of real-valued functions $f : Z \to \mathbb{R}$ on some domain $Z$. Given $z_1,\ldots,z_n \in Z$, we define the Rademacher complexity as follows:*

$$\mathcal{R}(F; z_{1,\ldots,n}) := \mathbb{E}\left[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}s_i f(z_i)\right],$$

*where the expectation is over $s_1,\ldots,s_n$ drawn uniformly at random from $\{-1,1\}^n$.*

The following theorem is a standard application of the Rademacher complexity for proving uniform convergence bounds.

**Proposition 50 (Uniform convergence from Rademacher complexity)** *Let $F$ be a family of functions $f : Z \to [a,b]$ on some domain $Z$ and with range bounded in $[a,b]$. Let $\Gamma$ be an arbitrary distribution over $Z$. Then for $n$ independent and identically distributed (i.i.d.) examples $z_1,\ldots,z_n$ from $\Gamma$,*

$$\mathbb{E}_{z_{1,\ldots,n}}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n}f(z_i) - \mathbb{E}_{z\sim\Gamma}[f(z)]\right|\right] \leq 2\,\mathbb{E}_{z_{1,\ldots,n}}[\mathcal{R}(F; z_{1,\ldots,n})].$$

*Moreover, for any $\delta \in (0, \frac{1}{3})$ and $n \in \mathbb{N}$, with probability at least $1 - \delta$ over the random draw of $n$ independent and identically distributed (i.i.d.) examples $z_1, \ldots, z_n$ from $\Gamma$, it holds that*

$$\sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} f(z_i) - \mathbb{E}_{z \sim \Gamma}[f(z)] \right| \leq 2\mathcal{R}(F; z_{1,\ldots,n}) + O\left( (b-a) \sqrt{\frac{\log(1/\delta)}{n}} \right).$$

**Proposition 51 (Rademacher Complexity after Lipschitz Postprocessing)** *Let $F$ be a family of functions $f : Z \to \mathbb{R}$. For $i = 1, \ldots, n$, let $z_i \in Z$ be an element of the domain $Z$ and let $\eta_i : \mathbb{R} \to \mathbb{R}$ be any 1-Lipschitz function. It holds that*

$$\mathbb{E}\left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} s_i \eta_i(f(z_i)) \right] \leq \mathcal{R}(F; z_{1,\ldots,n}) = \mathbb{E}\left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} s_i f(z_i) \right].$$

**Proof** By induction, it suffices to consider the case where all the $\eta_i$'s are the identity function except $\eta_1$.

Now we have

$$\begin{aligned}
&\mathbb{E}\left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} s_i \eta_i(f(z_i)) \right] \\
&= \frac{1}{2n} \mathbb{E}\left[ \sup_{f \in F} \left( s_1 \eta_1(f(z_1)) + \sum_{i=2}^{n} s_i f(z_i) \right) + \sup_{f \in F} \left( -s_1 \eta_1(f(z_1)) + \sum_{i=2}^{n} s_i f(z_i) \right) \right] \\
&= \frac{1}{2n} \mathbb{E}\left[ \sup_{f_+, f_- \in F} \left( \eta_1(f_+(z_1)) - \eta_1(f_-(z_1)) + \sum_{i=2}^{n} s_i (f_+(z_i) + f_-(z_i)) \right) \right] \\
&= \frac{1}{2n} \mathbb{E}\left[ \sup_{f_+, f_- \in F} \left( |\eta_1(f_+(z_1)) - \eta_1(f_-(z_1))| + \sum_{i=2}^{n} s_i (f_+(z_i) + f_-(z_i)) \right) \right]. \qquad (55)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{E}\left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} s_i f(z_i) \right] \\
&= \frac{1}{2n} \mathbb{E}\left[ \sup_{f_+, f_- \in F} \left( |f_+(z_1) - f_-(z_1)| + \sum_{i=2}^{n} s_i (f_+(z_i) + f_-(z_i)) \right) \right]. \qquad (56)
\end{aligned}$$

By the 1-Lipschitz property of $\eta_1$, we have

$$|\eta_1(f_+(z_1)) - \eta_1(f_-(z_1))| \leq |f_+(z_1) - f_-(z_1)|.$$

This implies that (55) is a lower bound of (56), completing the proof. ■

The following is the standard definition of the VC dimension for binary function families:

**Definition 52 (VC Dimension (Vapnik and Chervonenkis, 1971))** *The VC dimension of a family $F$ of binary functions $f : Z \to \{0, 1\}$ is the largest size $d$ of a subset $Z' = \{z_1, \ldots, z_d\} \subseteq Z$ such that for each of the $2^d$ choices of $\boldsymbol{s} := (s_1, \ldots, s_d) \in \{0, 1\}^n$, there exists $f_{\boldsymbol{s}} \in F$ such that $f_{\boldsymbol{s}}(z_i) = s_i$ for every $i = 1, \ldots, d$.*

The following standard result can be proved using Dudley's chaining argument (see e.g. Theorem 8.3.23 of Vershynin (2018)):

**Proposition 53 (Rademacher Complexity from VC Dimension)** *Let $F$ be a family of binary functions $f : Z \to \{0, 1\}$ with VC dimension at most $d$. Then for any $n \in \mathbb{Z}_{>0}$ and any $z_1, \ldots, z_n \in Z$, we have*

$$\mathcal{R}(F; z_{1,\ldots,n}) \leq O\left(\sqrt{\frac{d}{n}}\right).$$