A Unifying Theory of Distance from Calibration*

Jarosław Błasiok
Columbia University
New York City, USA
jb4451@columbia.edu

Parikshit Gopalan
Apple
Cupertino, USA
parikg@gmail.com

Lunjia Hu
Stanford University
Stanford, USA
lunjia@stanford.edu

Preetum Nakkiran
Apple
Cupertino, USA
preetum@nakkiran.org

ABSTRACT

We study the fundamental question of how to define and measure the distance from calibration for probabilistic predictors. While the notion of *perfect calibration* is well-understood, there is no consensus on how to quantify the distance from perfect calibration. Numerous calibration measures have been proposed in the literature, but it is unclear how they compare to each other, and many popular measures such as Expected Calibration Error (ECE) fail to satisfy basic properties like continuity.

We present a rigorous framework for analyzing calibration measures, inspired by the literature on property testing. We propose a ground-truth notion of distance from calibration: the ℓ_1 distance to the nearest perfectly calibrated predictor. We define a *consistent calibration measure* as one that is polynomially related to this distance. Applying our framework, we identify three calibration measures that are *consistent* and can be estimated efficiently: smooth calibration, interval calibration, and Laplace kernel calibration. The former two give quadratic approximations to the ground truth distance, which we show is information-theoretically optimal in a natural model for measuring calibration which we term the *prediction-only access* model. Our work thus establishes fundamental lower and upper bounds on measuring the distance to calibration, and also provides theoretical justification for preferring certain metrics (like Laplace kernel calibration) in practice.

CCS CONCEPTS

• **Mathematics of computing** → **Probability and statistics**; • **Theory of computation** → **Machine learning theory**.

KEYWORDS

probabilistic prediction, calibration, Kantorovich-Rubinstein duality

*We refer the readers to the full version of this paper [5].

STOC '23, June 20–23, 2023, Orlando, FL, USA

ACM ISBN 978-1-4503-9913-5/23/06…$15.00
https://doi.org/10.1145/3564246.3585182

ACM Reference Format:
Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. 2023. A Unifying Theory of Distance from Calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC '23), June 20–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3564246.3585182

1 INTRODUCTION

Probabilistic predictions are central to many domains which involve categorical, even deterministic, outcomes. Whether it is doctor predicting a certain incidence probability of heart disease, a meteorologist predicting a certain chance of rain, or an autonomous vehicle system predicting a probability of road obstruction— probabilistic prediction allows the predictor to incorporate and convey epistemic and aleatory uncertainty in their predictions.

In order for predicted probabilities to be operationally meaningful, and not just arbitrary numbers, they must be accompanied by some form of formal probabilistic guarantee. The most basic requirement of this form is *calibration* [8]. Given a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$ representing points with binary labels, a predictor $f : \mathcal{X} \to [0,1]$ which predicts the probability of the label being 1 is calibrated if for every $v \in \mathrm{Im}(f)$, we have $\mathbb{E}[y|f(x) = v] = v$. Calibration requires that, for example, among the set of patients which are predicted to have a 10% incidence of heart disease, the true incidence of heart disease is exactly 10%. Calibration is recognized as a crucial aspect of probabilistic predictions in many applications, from their original development in meteorological forecasting [11, 22, 37, 38], to models of risk and diagnosis in medicine [6, 12, 24, 30, 34, 35, 49], to image classification settings in computer vision [36, 39]. There is a large body of theoretical work on it in forecasting, for example [10, 14, 15, 25]. More recently, the work of [23] on multicalibration as a notion of group fairness (see also [27, 29]), and connections to indistinguishability [13] and loss minimization [17, 18] have spurred renewed interest in calibration from theoretical computer science.

In practice, it is of course rare to encounter *perfectly* calibrated predictors, and thus it is important to quantify their *distance from calibration*. However, while there is consensus across domains on what it means for a predictor to be *perfectly calibrated*, there is no consensus even within a domain on how to measure this distance. This is because the commonly-used metrics of calibration have fundamental theoretical flaws, which manifest as practical frustrations. Consider the Expected Calibration Error (ECE), which

is the de-facto standard metric in the machine learning community (e.g. [20, 36, 40, 44]).

Definition 1.1. For a predictor $f : \mathcal{X} \to [0,1]$ and distribution $\mathcal{D}$ over $(x, y) \in \mathcal{X} \times \{0,1\}$, the expected calibration error $\mathsf{ECE}_{\mathcal{D}}(f)$ is defined as

$$\mathsf{ECE}_{\mathcal{D}}(f) = \mathop{\mathbb{E}}_{\mathcal{D}}\left[\left|\mathop{\mathbb{E}}_{\mathcal{D}}[y \mid f(x)] - f(x)\right|\right].$$

The ECE has a couple of flaws: First, it is impossible to estimate in general from finite samples (e.g. [33, Proposition 5.1] and [1]). This is partly because estimating the conditional expectation $\mathbb{E}[y|f(x) = v]$ requires multiple examples with the exact same prediction $f(x) = v$, which could happen with arbitrarily low probability in a fixed and finite number of examples over a large domain $\mathcal{X}$. Second, the ECE is *discontinuous* as a function of the predictor f, as noted by [14, 25]. That is, arbitrarily small perturbations to the predictor f can cause large fluctuations in $\mathsf{ECE}_{\mathcal{D}}(f)$. We illustrate this with a simple example below.

Consider the uniform distribution over a two-point space $X = \{a, b\}$, with the label for a is 0 whereas for b it is 1. The predictor $\overline{f}$ which always predicts $1/2$ is perfectly calibrated under $\mathcal{D}$, so $\mathsf{ECE}(\overline{f}) = 0$. In contrast, the related predictor where $f(a) = (1/2 - \varepsilon)$ and $f(b) = (1/2 + \varepsilon)$, for arbitrarily small $\varepsilon > 0$, has $\mathsf{ECE}(f) = 1/2 - \varepsilon$. Thus the infinitesimal change from $\overline{f}$ to f causes a jump of almost $1/2$ in ECE.

This discontinuity also presents a barrier to popular heuristics for estimating the ECE. For example, the estimation problem is usually handled by discretizing the range of f, yielding an alternate quantity — the "binned-ECE"— that can be estimated from samples [41]. However, the choice of discretization turns out to matter significantly both in theory [31, Example 3.2] and in practice [42]. For example, both [36, Section 5] and [42, Section 5.3] found that changing the number of bins used in binned-ECE can change conclusions about which of two models is better calibrated. In the simple two-point example above, if we choose m bins of equal width, then we observe a binned-ECE of either 0 or $\approx 1/2$, depending on whether m is odd or even!

To address the shortcomings of the ECE, a long line of works have proposed alternate metrics of miscalibration. These metrics take a diversity of forms: some are based on modifications to the ECE (e.g. alternate binning schemes, debiased estimators, or smoothing) [26, 31, 33, 46, 50], some use proper scoring rules, some rely on distributional tests such as Kolmogorov–Smirnov [21], Kernel MMD [32], or other nonparametric tests [1]. Yet it is not clear what to make of this smorgasbord of calibration metrics: whether these different metrics are at all related to each other, and whether they satisfy desirable properties (such as continuity). For a practitioner training a model, if their model is calibrated under some of these notions, but not others, what are they to make of it? Should they report the most optimistic metrics, or should they strive to be calibrated for all of them? Or is there some inherent but undiscovered reason why all these metrics should paint a similar picture?

Underlying this confusion is a foundational question: *what is the ground truth distance of a predictor from calibration?* To our knowledge, this question has not been answered or even asked in the prior literature. Without a clearly articulated ground truth and a set of desiderata that a calibration measure must satisfy, we cannot hope to have meaningful comparisons among metrics.

At best one can say that ECE and (certain but not all) binning based variants give an upper bound on the true distance to calibration; we prove this formally for ECE in section 4.3. Thus if a predictor can be guaranteed to have small ECE, then it is indeed close to being calibrated in a formal sense (see for instance [23, Claim 2.10]). But small ECE might an unnecessarily strong (or even impossible) constraint to satisfy in many realistic settings, especially when dealing with predictors which are allowed to produce real-valued outputs. For example, consider the standard setting of a deep neural network trained from random initialization for binary classification. The predicted value $f(x) \in [0,1]$ is likely to be different for every individual x in the population, which could result in a similar situation to our example. The ECE is likely to greatly overstate the true distance from calibration in such a setting.

This brings us to the main motivations behind this work. We aim to:

- Formulate desiderata for good calibration measures, based on a rigorous notion of ground truth distance from calibration.
- Use our desiderata to compare existing calibration measures, identifying measures that are good approximations to the ground truth distance.
- Apply theoretical insights to inform practical measurements of calibration in machine learning, addressing known shortcomings of existing methods.

Summary of Our Contributions. We summarize the main contributions of our work:

- **Framework for measuring the distance to calibration (Section 4).** We propose a ground truth notion of distance to calibration which is the ℓ_1 distance to the closest perfectly calibrated predictor, inspired by the property testing literature [3, 16]. We define the set of *consistent calibration measures* to be those that provide polynomial upper and lower bounds on the true distance.
- **Consistent calibration measures.** We identify three calibration measures that are in fact consistent: two have been proposed previously [25, 32] and the third is new. Interestingly, the two prior measures (smooth and kernel calibration) were proposed with other motivations in mind, and not as standalone calibration measures. We consider it surprising that they turn out to be intimately related to the ground truth ℓ_1 distance.
  (1) **Interval calibration error (Section 6)**. This is a new measure which is reminiscent of the binning estimate that is popular in practice [20, 36, 40]. We show that by randomizing both the width of each bin and using a random offset, and by adding the average bin width to the resulting calibration error, one can derive a consistent estimator that this is always an upper bound on the true distance, and it is never more than the square root of the true distance.
  (2) **Smooth calibration error (Section 7)** was proposed in the work of [25]. We show using LP duality that it is a constant factor approximation of the *lower* distance to calibration, which we define to be, roughly speaking, a particular Wasserstein distance to perfect calibration. The

lower distance to calibration is always at most the true distance to calibration and is always at least a constant times the true distance squared.

(3) **Laplace-kernel calibration error (Section 8).** This is a calibration measure that was proposed in the work of [32]. While they did not recommend a particular choice of kernel, we show that using the Laplace kernel happens to yield a consistent measure, while using the Gaussian kernel does not.

In contrast to these measures, other commonly used heuristics (ECE and binning based) do not meet our criteria for being consistent calibration measures. Our work thus provides a firm theoretical foundation on which to base evaluations and comparisons of various calibration measures; such a foundation was arguably lacking in the literature.

- **Matching lower bounds.** Smooth calibration and interva calibration provide quadratic approximations to the true distance from calibration. We prove that this is the best possible approximation, by showing an information-theoretic barrier: for calibration measures depending only on the labels y and predictions f, which are oblivious to the points x themselves (as most calibration measures are), it is impossible to obtain better than a quadratic approximation to the true distance from calibration. Thus, the measures above are in fact optimal in this sense.
- **Better efficiency in samples and run time.** We present improved algorithms and sample complexity bounds for computing some calibration measures. We present the first efficient algorithm for computing smooth calibration error using a linear program. We also observe that the techniques of [45] yield an alternate algorithm to computing kernel calibration error which is (somewhat surprisingly) reminiscent of randomized binning.
- **Insights for Practice.** Our results point to concrete takeaways for practical measurements of calibration. First, we recommend using either Laplace kernel calibration or Interval calibration, as calibration measures that are theoretically consistent, computationally efficient, and simple to implement. Second, if Binned-ECE must be used, we recommend randomly shifting the bin boundaries together, and adding the average width of the bins to the calibration estimate. These modifications turn binning into a upper-bound on calibration distance, and bring it closer to interval calibration error which is a consistent calibration measure (Section 2.3). Finally, we experimentally evaluate our calibration measures on a family of synthetic data distributions, to demonstrate their behavior in more natural settings (beyond worst-case guarantees) (see Section 10 of the full version [5]).

Organization of this paper. The rest of this paper is organized as follows. In Section 2 we present an informal overview of our main results, highlighting the definitions and key conceptual ideas. We discuss related works in Section 3. Section 4 sets up our notion of true distance from calibration and the desiderata that we seek from calibration measures. We also explain how ECE and some other measures fail these desiderata. Section 5 defines the upper and lower distance to calibration. Section 6 analyzes Interval Calibration error, Section 7 analyzes Smooth calibration error and Section 8 analyzes the Laplace kernel calibration error. In the full version of this paper [5] we give sample complexity bounds and efficient algorithms for estimating various calibration measures using random sample, and we experimentally evaluate our calibration measures on a representative family of synthetic data distributions. The full version [5] includes all the technical proofs for this paper.

2 OVERVIEW OF OUR RESULTS

We start by setting up some notation for calibration in the binary classification setting. Let $\mathcal{X}$ be a discrete domain, defining the input space. We are given samples (x, y) drawn from a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$. A *predictor* is a function $f : \mathcal{X} \to [0, 1]$, where $f(x)$ is interpreted as an estimate of $\Pr[y = 1 \mid x]$. For a predictor f and distribution $\mathcal{D}$, we often consider the induced joint distribution of prediction-label pairs $(f(x), y) \in [0, 1] \times \{0, 1\}$, which we denote $\mathcal{D}_f$. We say a prediction-label distribution Γ over $[0, 1] \times \{0, 1\}$ is *perfectly calibrated* if $\mathbb{E}_{(v,y)\sim\Gamma}[y \mid v] = v$. For a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$, we say a predictor $f : \mathcal{X} \to [0, 1]$ is *perfectly calibrated w.r.t.* $\mathcal{D}$ if the induced distribution $\mathcal{D}_f$ is perfectly calibrated. Finally, a *calibration measure* μ is a function that maps a distribution $\mathcal{D}$ and a predictor $f : \mathcal{X} \to [0, 1]$ to a value $\mu_{\mathcal{D}}(f) \in [0, 1]$.

2.1 Framework for Measuring Distance from Calibration

The primary conceptual contribution of this work is a formal framework in which we can reason about and compare various measures of calibration. We elaborate upon the key ingredients of this framework.

The true distance to calibration. We define the ground truth distance from calibration as the distance to the closest calibrated predictor. Measuring distance requires a metric on the space of all predictors. A natural metric is the ℓ_1 metric given by $\ell_1(f, g) = \mathbb{E}_{\mathcal{D}} |f(x) - g(x)|$. Accordingly we define the true distance from calibration as

$$\mathrm{dCE}_{\mathcal{D}}(f) := \inf_{g \in \mathrm{cal}(\mathcal{D})} \mathop{\mathbb{E}}_{\mathcal{D}} |f(x) - g(x)|, \tag{1}$$

where $\mathrm{cal}(\mathcal{D})$ denotes the set of predictors that are perfectly calibrated w.r.t. $\mathcal{D}$. This definition is intuitive, and natural from a property testing point of view [3, 16, 43], but has not been proposed before to our knowledge. Note that it is not clear how to compute this distance efficiently: the set $\mathrm{cal}(\mathcal{D})$ is non-convex, and in fact it is discrete when the domain $\mathcal{X}$ is discrete. A more subtle issue is that it depends on knowing the domain $\mathcal{X}$, whereas traditionally calibration measures only depend on the joint distribution $\mathcal{D}_f$ of predictions and labels.

Access model. Calibration measures $\mu_{\mathcal{D}}(f)$ can depend on the entire distribution $\mathcal{D}$, as well as on the predictor $f : \mathcal{X} \to [0, 1]$. However, we would prefer measures which only depend on the prediction-label joint distribution $\mathcal{D}_f$, similar to standard loss functions in machine learning and classic calibration measures [9, 10, 15]. This distinction has important consequences for the power of calibration measures, which we describe shortly. We delineate the two levels of access as follows:

(1) **Sample access (SA).** In the SA model, $\mu_{\mathcal{D}}(f)$ is allowed to depend on the full joint distribution $(x, f(x), y)$ for $(x, y) \sim \mathcal{D}$. This terminology follows [13].
(2) **Prediction-only access (PA).** In the PA model, $\mu_{\mathcal{D}}(f)$ is only allowed to depend on $\mathcal{D}_f$, the joint distribution $(f(x), y)$ for $(x, y) \sim \mathcal{D}$. In particular, μ cannot depend on the input domain $\mathcal{X}$.

Observe that the ground truth distance (dCE) is defined in the sample access model, since Equation (1) depends on the domain and distribution of x. On the other hand, we often desire measures that can be computed in the prediction access model.

Robust completeness and soundness. We propose two desiderata for any calibration measure μ: robust completeness and robust soundness, in analogy to completeness and soundness in proof systems.

(1) **Robust completeness** requires $\mu_{\mathcal{D}}(f) \le O(\mathsf{dCE}_{\mathcal{D}}(f)^c)$ for some constant c. This guarantees that any predictor which is close to a perfectly calibrated predictor (in ℓ_1) has small calibration error under μ. This is a more robust guarantee than standard completeness, which in this setting would mean just that $\mathsf{dCE}_{\mathcal{D}}(f) = 0$ implies $\mu_{\mathcal{D}}(f) = 0$, but would not give any guarantees when $\mathsf{dCE}_{\mathcal{D}}(f)$ is non-zero but small.
(2) **Robust soundness** requires $\mu_{\mathcal{D}}(f) \ge \Omega(\mathsf{dCE}_{\mathcal{D}}(f)^s)$ for some constant s. That is, if $\mathsf{dCE}_{\mathcal{D}}(f)$ is large then so is $\mu_{\mathcal{D}}(f)$.

We call a calibration measure *consistent* (or more precisely, (c, s)-*consistent*) if it satisfies both robust completeness and robust soundness, for some parameters $c, s > 0$:

$$\Omega(\mathsf{dCE}_{\mathcal{D}}(f)^s) \le \mu_{\mathcal{D}}(f) \le O(\mathsf{dCE}_{\mathcal{D}}(f)^c). \tag{2}$$

Consistent measures are exactly those that are polynomially-related to the true distance from calibration, $\mathsf{dCE}_{\mathcal{D}}(f)$. The reader might wonder if, in our definition of consistent calibration measures (Equation 2), we could require *constant factor* approximations to $\mathsf{dCE}_{\mathcal{D}}(f)$ rather than polynomial factors. It turns out that there are information-theoretic barriers to such approximations in the prediction-access model. The core obstacle is that the true distance dCE is defined in the SA model, and one cannot compute it exactly in the prediction-only access model or approximate it within a constant factor. Indeed, we show that any calibration measure computable in the prediction-access must satisfy $s/c \ge 2$: information theoretically, a quadratic approximation is the best possible.

Another nice property of our definition is that the set of all consistent measures stays the same, even if we define distances between predictors using the ℓ_p metric for $p > 1$ in place of ℓ_1, since all ℓ_p measures are polynomially related.

Desiderata for calibration measures. Given the discussion so far, we can now stipulate three desiderata that we would like calibration measures μ to satisfy:

(1) **Access:** μ is well-defined in the Prediction-only access model (PA).
(2) **Consistency:** μ is (c, s)-consistent – that is, μ is polynomially related to the true distance from calibration dCE. Ideally we have $s/c = 2$, which is optimal in the PA model (Corollary 4.6).
(3) **Efficiency:** $\mu_{\mathcal{D}}(f)$ can be computed within accuracy ε in time $\mathrm{poly}(1/\varepsilon)$ using $\mathrm{poly}(1/\varepsilon)$ random samples from $\mathcal{D}_f$.

Various notions that have been proposed in the literature fail one or more of these desiderata; ECE for instance fails robust completeness since an arbitrarily small perturbation of a perfectly calibrated predictor could result in high ECE. We refer the reader to Table 1 for a more complete treatment of such notions.

2.2 Information-Theoretic Limitations of the Prediction-Access Model

Upper and Lower Distances. We start with the following question, which formalizes how well one can approximate the true distance to calibration in the prediction-only access (PA) model.

For a given distribution $\mathcal{D}$ and predictor f, how large or small can $\mathsf{dCE}_{\mathcal{D}'}(f')$ be, among all other $(\mathcal{D}', f')$ which have the same prediction-label distribution ($\mathcal{D}'_{f'} = \mathcal{D}_f$)?

We denote the minimum and maximum using $\underline{\mathsf{dCE}}_{\mathcal{D}}(f)$ and $\overline{\mathsf{dCE}}_{\mathcal{D}}(f)$ respectively, which we call the lower and upper distance to calibration respectively. Hence

$$\underline{\mathsf{dCE}}_{\mathcal{D}}(f) \le \mathsf{dCE}_{\mathcal{D}}(f) \le \overline{\mathsf{dCE}}_{\mathcal{D}}(f). \tag{3}$$

Both these quantities are defined in the PA model, in which they represent the tightest lower and upper bounds respectively that one can prove on dCE. As framed, they involve considering all possible domains and distributions $\mathcal{D}'$ over them. But we can give simpler characterizations of these notions.

The upper distance $\overline{\mathsf{dCE}}$ can be alternatively viewed as the minimum distance to calibration via post-processing: it is the distance to the closest calibrated predictor $g \in \mathsf{cal}(\mathcal{D})$ such that $g = \kappa(f)$ can be obtained from f by *post-processing* its predictions. For the lower distance $\underline{\mathsf{dCE}}$, we can abstract away the domain and ask only for a coupling between f and a perfectly calibrated predictor:

Consider all joint distributions Π of (u, v, y) over $[0, 1] \times [0, 1] \times \{0, 1\}$ where $(v, y) \sim \mathcal{D}_f$ and the distribution of (u, y) is perfectly calibrated. How small can $\mathbb{E}\,|u - v|$ be?

Limiting ourselves to couplings of the form $(g(x), f(x), y) \sim \mathcal{D}$ where $g \in \mathsf{cal}(\mathcal{D})$ would recover dCE. Our definition also permits couplings that may not be realizable on the domain $\mathcal{X}$, giving a lower bound.

An equivalent view of these distances is that the upper distance only considers those calibrated predictors whose level sets are obtained by a coarsening of the level sets of f. The lower distance allows calibrated predictors that are obtained by a finer partitioning of the level sets of f. Theorem 5.5 proves the equivalence of these various formulations of $\underline{\mathsf{dCE}}$ and $\overline{\mathsf{dCE}}$.

A Quadratic Barrier. How tight are the lower and upper bounds in Equation (3)? That is, how tightly can dCE be determined in the Prediction-only access model? In Lemma 4.5, we show that there can be at least a quadratic gap in between any two adjacent terms in Equation (3). We construct two distributions $\mathcal{D}^1$ and $\mathcal{D}^2$ and a predictor f such that

- $\mathcal{D}^1_f = \mathcal{D}^2_f$, so the upper and the lower distance are equal for both distributions, but they are well-separated from each

other; $\underline{\mathsf{dCE}}_{\mathcal{D}^i}(f) = \Theta(\alpha^2)$ whereas $\overline{\mathsf{dCE}}_{\mathcal{D}^i}(f) = \Theta(\alpha)$ for $i \in \{1, 2\}$.
- $\mathsf{dCE}_{\mathcal{D}^i}(f)$ equals either $\underline{\mathsf{dCE}}$ or $\overline{\mathsf{dCE}}$ depending on whether $i = 1$ or 2.

This example raises the question of whether an even bigger gap can exist, which we answer next.

2.3 Consistent Calibration Measures

We describe three consistent calibration measures, and their relation to the true distance from calibration.

Interval calibration (Section 6). Interval calibration error is a subtle modification to the heuristic of binning predictions into buckets and computing the expected calibration error. Formally, given a partition $\mathcal{I} = \{I_1, \ldots, I_m\}$ of $[0, 1]$ into intervals of width bounded by $w(\mathcal{I})$, we first consider the standard quantity

$$\mathsf{binnedECE}_{\mathcal{D}}(f, \mathcal{I}) = \sum_{j \in [m]} |\mathbb{E}[(f - y)\mathbf{1}(f \in I_j)]|. \tag{4}$$

This quantity, as the name suggests, is exactly the Binned-ECE for the bins defined by the partition $\mathcal{I}$. We then define our notion of *Interval calibration error* (intCE) as the minimum of this Binned-ECE over all partitions $\mathcal{I}$, when "regularized" by maximum bin width $w(\mathcal{I})$:

$$\mathsf{intCE}_{\mathcal{D}}(f) := \inf_{\mathcal{I}\text{: Interval partition}} \left(\mathsf{binnedECE}_{\mathcal{D}}(f, \mathcal{I}) + w(\mathcal{I})\right).$$

In Theorem 6.2, we show that intCE satisfies the following bounds.

$$\overline{\mathsf{dCE}}_{\mathcal{D}}(f) \le \mathsf{intCE}_{\mathcal{D}}(f) \le 4\sqrt{\underline{\mathsf{dCE}}_{\mathcal{D}}(f)}.$$

This shows that the measure intCE is $(1/2, 1)$-consistent, and gives the best possible (quadratic) approximation to the true distance to calibration. The outer inequality implies that the gap between the lower and upper distance is no more than quadratic, hence the gap exhibited in Lemma 4.5 is tight.

We now address the computational complexity. While the definition of interval calibration minimizes over all possible interval partitions, in Section 6.2 of the full version [5], we show that it suffices to consider a geometrically decreasing set of values for the width w, with a random shift, to get the desired upper bound on dCE.

Our result suggests an additional practical takeaway: if the standard binning algorithm must be used to measure calibration, then the bin width should be added to the binnedECE. This yields a quantity which is at least an *upper bound* on the true distance to calibration, which is not true without adding the bin widths. Specifically, for *any* interval partition $\mathcal{I}$, we have:

$$\overline{\mathsf{dCE}}_{\mathcal{D}}(f) \le \mathsf{binnedECE}(f, \mathcal{I}) + w(\mathcal{I}). \tag{5}$$

Thus, if we add the bin width, then binnedECE can at least be used to certify closeness to calibration. The extreme case of width 0 buckets corresponds to ECE, while the case when the bucket has width 1 corresponds to the weaker condition of accuracy in expectation [23]. It is natural to penalize larger width buckets which allow cancellations between calibration errors for widely separated values of f. The notion of using bucket width as a penalty to compare calibration error results obtained from using differing width buckets is intuitive in hindsight, but not done in prior work to our knowledge (e.g. [36]).

Smooth Calibration (Section 7). Smooth calibration is a calibration measure first defined by [25], see also [14, 19]. Smooth calibration error is defined as the following maximization over the family L of all bounded 1-Lipschitz functions $w : [0, 1] \to [-1, 1]$:

$$\mathsf{smCE}_{\mathcal{D}}(f) := \mathsf{smCE}(\mathcal{D}_f) = \sup_{w \in L} \mathop{\mathbb{E}}_{(v,y)\sim\mathcal{D}_f} [w(v)(y - v)].$$

Without the Lipschitz condition on functions w, this definition would be equivalent to $\mathsf{ECE}(f)$. Adding the Lipschitz condition smooths out the contribution from each neighborhood of v and results in a calibration measure that is Lipschitz in f with respect to the ℓ_1 distance. This notion has found applications in game theory and leaky forecasting [14, 25]. Our main result is that the smooth calibration error captures the lower distance from calibration up to constant factors:

$$\frac{1}{2}\,\underline{\mathsf{dCE}}_{\mathcal{D}}(f) \le \mathsf{smCE}_{\mathcal{D}}(f) \le 2\,\underline{\mathsf{dCE}}_{\mathcal{D}}(f).$$

We find this tight connection to be somewhat surprising, since smCE (as a maximization over weight functions w) and $\underline{\mathsf{dCE}}$ (as a minimization over couplings) have *a priori* very different definitions. They turn out to be related via LP duality, in a way analogous to Kantorovich-Rubinstein duality of Wasserstein distances. We present a high-level overview of the proof at the start of Section 7. Along the way, we give an efficient polynomial time algorithm for estimating the smooth calibration error, the first such algorithm to our knowledge. To summarize the relations between notions discussed so far, we have

$$\boxed{\mathsf{smCE} \approx \underline{\mathsf{dCE}} \le \mathsf{dCE} \le \overline{\mathsf{dCE}} \le \mathsf{intCE} \le 4\sqrt{\underline{\mathsf{dCE}}}} \tag{6}$$

For each of the first three inequalities, we show that the gap can be quadratic. The final inequality shows that these gaps are *at most* quadratic.

Kernel Calibration (Section 8). The notion of kernel calibration error was introduced in [32] as *Maximum Mean Calibration Error* (MMCE). Kernel calibration can be viewed as a variant of smooth calibration error, where we use as weight functions $w : [0, 1] \to \mathbb{R}$ which are bounded with respect to a norm $\|\cdot\|_K$ on the *Reproducing Kernel Hilbert Space* associated with some positive-definite kernel K:

$$\mathsf{kCE}^K_{\mathcal{D}}(f) := \sup_{w : \|w\|_K \le 1} \mathop{\mathbb{E}}_{(v,y)\sim\mathcal{D}_f} [w(v)(y - v)].$$

When $\mathcal{D}_f$ is an empirical distribution over samples $\{(v_1, y_1), \ldots, (v_n, y_n)\}$, this can be computed as

$$\mathsf{kCE}^K_{\mathcal{D}}(f) = \sqrt{\frac{1}{n^2}\sum_{i,j}(y_i - v_i)(y_j - v_j)K(v_i, v_j)}.$$

The original motivation of introducing the kernel calibration error was to provide a differentiable proxy for ECE — allowing for the calibration error to be explicitly penalized during the training of a neural network. However, [32] does not discuss how the choice of the kernel affects the resulting measure, although they used Laplace kernel in their experiments. We prove here that this choice has strong theoretical justification — the kernel calibration error

Table 1: Calibration measures proposed in, or based on, prior works. Here P.S.R. is short for proper scoring rules.

Metric	Continuity	Completeness	Soundness
(ℓ_p-)ECE	✗	✓	✓
Binned-ECE	✗	✓	✗
P.S.R. (Brier, NLL)	✓	✗	✓
NCE [48]	✓	✗	✓
ECCE [1]	✗	✓	✓
MMCE [32]	✓	✓	✓
smCE [25]	✓	✓	✓

with respect to Laplace kernel is a consistent calibration measure; specifically for some positive absolute constants $c_1, c_2 > 0$,

$$c_1 \underline{\mathsf{dCE}}(f) \le \mathsf{kCE}^{\mathsf{Lap}}(f) \le c_2 \sqrt{\underline{\mathsf{dCE}}(f)}.$$

This says that we can view kernel calibration with respect to the Laplace kernel as fundamental measure in its own right, as opposed to a proxy for (the otherwise flawed) ECE. We also show that the choice of kernel is in fact crucial: for the Gaussian kernel, another commonly used kernel across machine learning, the resulting measure is not robustly sound anymore (Theorem 8.6).

2.4 Better Algorithms and Sample Complexity

For many of the measures discussed in the paper, we provide efficient algorithms yielding an ε additive approximation to the measure in question, using samples from the distribution $\mathcal{D}_f$. In most cases, those results follow a two step paradigm, we give an algorithm that approximates the measure on a finite sample, followed by a generalization bound. Our generalization bounds follow from essentially standard bounds on Rademacher complexity of the function families involved in defining our measures (e.g. bounding the Rademacher complexity of 1-Lipshitz functions for smCE). On the algorithmic side, we prove that the smCE on the empirical distribution over a sample of size n can be computed by solving a linear program with $O(n)$ variables and constraints. Similarly, the $\underline{\mathsf{dCE}}$ can be approximated up to an error ε, by linear time prepossessing followed by solving a linear program with $O(\varepsilon^{-1})$ variables and constrains.

We provide an alternate algorithm for estimating the kernel calibration error with Laplace kernel, using the *Random Features Sampling* technique from [45]. This algorithm does not improve on naive estimators in worst-case guarantees, but it reveals an intriguing connection. After unwrapping the random features abstraction, the final algorithm is similar to the popular interval binning calibration estimator, where we choose the length of the interval at random from a specific distribution, and introduce a uniformly random shift. We find it surprising that an estimate of this type is *exactly* equal to $(\mathsf{kCE}^{\mathsf{Lap}})^2$ in expectation.

3 RELATED WORK

We start by discussing the high-level relation between our work and other areas of theoretical computer science, and then discuss work on calibration in machine learning and forecasting.

Property testing. Our framework for defining the distance to calibration is inspired by the elegant body of literature on property testing [3, 16, 47]. Indeed, the notions of ground truth distance to calibration, robust completeness and robust soundness are in direct correspondence to notions in the literature on tolerant property testing and distance estimation [43]. Like in property testing, algorithms for estimating calibration measures operate under stringent resource constraints, although the constraints are different. In property testing, the algorithm only has a local view of the object based on a few queries. In our setting, the constraint comes from having the operate in the prediction-only access model whereas the ground truth distance is defined in the sample-access model.

Multicalibration. Recent interest in calibration and its variants in theoretical computer science has been spurred by the work of [23] introducing multicalibration as a group fairness notion (see also [28, 29]). This notion has proved to be unexpectedly rich even beyond the context of fairness, with connections to indistinguishability [13] and loss minimization [18]. Motivated by the goal of finding more computationally efficient notions of multicalibration, notions such as low-degree multicalibration [19] and calibrated multiaccuracy have been analyzed in the literature [17], some of these propose new calibration measures.

Level of access to the distribution. The sample access model is considered in the work of [13], who relate it to the notion of multicalibration [23]. Prediction-only access is a restriction of sample access which is natural in the context of calibration, and is incomparable to the no-access model of [13] where on gets access to point label pairs. This model is not considered explicitly in [13], and the name prediction-only access for it is new. But the model itself is well-studied in the literature on calibration [7, 10, 15], indeed all existing notions of calibration that we are aware of are defined in the PA model, as are the commonly used losses in machine learning.

Prior work on calibration measures. Several prior works have proposed alternate measures of calibration (Table 1 lists a few of them). Most focus on the *how*: they give a formula or procedure for computing of the calibration measure from a finite set of samples, sometimes accompanied by a generalization guarantee that connects it to some property of the population. There is typically not much justification for *why* the population quantity is a good measure of calibration error, or discussion of its merits relative to other notions in the literature (notable exceptions are the works of [14, 25]). The key distinction in our work is that we start from a clear and intuitive ground truth notion and desiderata for calibration measures, we analyze measures based on how well they satisfy these desiderata, and then give efficient estimators and generalization guarantees for consistent calibration measures.

Our desiderata reveal important distinctions between measures that were proposed previously; Table 1 summarizes how well calibration measures suggested in prior works satisfy our desiderata. It shows that a host of calibration measures based on variants of ECE, binning and proper scoring rules fail to give basic guarantees. We present these guarantees formally in section 4.3. Briefly, many ECE variants suffer from the same flaws as ECE itself, and proper scoring rules suffer different issues we describe below. More discussions

about other notions of calibration from the literature can be found in Appendix A of the full version [5].

Proper Scoring Rules. Proper scoring rules such as the Brier Score [4] or Negative-Log-Loss (NLL) are popular proxies for miscalibration. Every proper scoring rule satisfies soundness, since if the score is 0, the function f is perfectly calibrated. However, such rules violate completeness: there are perfectly-calibrated functions for which the score is non-zero. For example, if the true distribution on labels $p(y|x) = \text{Bernoilli}(0.5)$, then the constant function $f(x) = 0.5$ is perfectly calibrated but has non-zero Brier score. This is because proper scoring rules measure predictive quality, not just calibration. The same holds for Normalized Cross Entropy (NCE), which is sound but not complete.

Smooth and Kernel Calibration. We show that some definitions in the literature do satisfy our desiderata— namely the notions of *weak calibration* (smooth calibration in our terminology) introduced in [25], and *MMCE* (calibration with a Laplace kernel in our terminology) introduced in [32]. Smooth calibration was introduced under the name "weak calibration" in [25], the terminology of smooth calibration is from [19][1]. Interestingly, these were developed with different motivations. MMCE was proposed by [32] for practical reasons: as a differentiable proxy for ECE, to allow optimizing for calibration via backpropagation. One of the motivations behind smooth calibration, discussed in both [14, 25] was to address the discontinuity of the standard binning measures of calibration and ECE. But it main application was as a weakening of perfect calibration, to study the power of deterministic forecasts in the online setting and derandomize the classical result of [15] on calibrated forecasters.

Our work establishes that these measures are not just good ways to measure calibration, they are more fundamental than previously known. Smooth calibration is within constant factors of the lower distance to calibration, and yields the best possible quadratic approximation to the true distance to calibration.

4 A FRAMEWORK FOR CALIBRATION MEASURES

In this section, we will present our framework for calibration measures. We start by characterizing the set of perfectly calibrated predictors. We then propose our ground truth notion of distance from calibration, in analogy to the distance from a code in property testing. Building on this, we formulate robust completeness and soundness guarantees that we want calibration measures to satisfy. Finally, we show information theoretic reasons why any calibration measure can only hope to give a quadratic approximation to the ground truth distance. We start with some notation.

Notation. Let $\mathcal{X}$ be a discrete domain.[2] Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{0,1\}$; we denote a sample from $\mathcal{D}$ by $(x,y) \sim \mathcal{D}$ where $x \in \mathcal{X}, y \in \{0,1\}$. A predictor is a function $f : \mathcal{X} \to [0,1]$, where $f(x)$ is an estimate of $\Pr[y = 1|x]$. We define the Bayes optimal predictor f^* as $f^*(x) = \mathbb{E}[y|x]$. Note that $\mathcal{D}$ is completely specified by the marginal distribution $\mathcal{D}_\mathcal{X}$ on $\mathcal{X}$, and the conditional expectations f^*. We let $\mathcal{F}_\mathcal{X}$ denote the set of all predictors $f : \mathcal{X} \to [0,1]$. We define the ℓ_1 distance in $\mathcal{F}_\mathcal{X}$ as

$$\ell_1(f,g) = \mathop{\mathbb{E}}_{\mathcal{D}} |f(x) - g(x)|.$$

For a distribution $\mathcal{D}$ and predictor f, we use $\mathcal{D}_f$ to denote the distribution over $\text{Im}(f) \times \{0,1\}$ of $(f(x), y)$ where $(x,y) \sim \mathcal{D}$. Two predictors f and g might be far apart in ℓ_1, yet D_f and $\mathcal{D}_g$ can be identical. [3]

A calibration measure μ is a function that maps a distribution $\mathcal{D}$ and a predictor $f : \mathcal{X} \to [0,1]$ to a value $\mu_\mathcal{D}(f) \in [0,1]$. A crucial question is the level of access to the underlying distribution that a procedure for computing μ has. We refer to the setting where an algorithm has access to $(x, f(x), y)$ for $(x,y) \sim \mathcal{D}$ as the sample-access model or SA model for short following [13]. Calibration measures are typically defined in the more restricted *prediction-only access model* or PA model for short, where we only get access to the joint distribution $\mathcal{D}_f$ of prediction-label pairs (f,y). Such calibration measures μ can be defined as follows: we first define $\mu(\Gamma) \in [0,1]$ for every distribution Γ over $[0,1] \times \{0,1\}$, and then for a distribution $\mathcal{D}$ and a predictor f, we define $\mu_\mathcal{D}(f)$ to be $\mu(\mathcal{D}_f)$.

We say a distribution Γ over $[0,1] \times \{0,1\}$ is *perfectly calibrated* if $\mathbb{E}_{(v,y)\sim\Gamma}[y|v] = v$. For a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, we say a predictor $f : \mathcal{X} \to [0,1]$ is *perfectly calibrated* w.r.t. $\mathcal{D}$ if $\mathcal{D}_f$ is perfectly calibrated. We use $\text{cal}(\mathcal{D})$ to denote the set of predictors f that is perfectly calibrated w.r.t. $\mathcal{D}$.

There is an injection from $\text{cal}(\mathcal{D})$ to the set of partitions of the domain $\mathcal{X}$. A consequence is that when $\mathcal{X}$ is finite, so is $\text{cal}(\mathcal{D})$. In particular, $\text{cal}(\mathcal{D})$ is not a convex subset of $\mathcal{F}_\mathcal{X}$. We describe the injection below for completeness, although it is not crucial for our results. For a partition $\mathcal{S} = \{S_i\}_{i=1}^m$, we define $g_\mathcal{S}(x) = E[y|x \in S_i]$ for all $x \in S_i$. It is clear that $g_\mathcal{S} \in \text{cal}(\mathcal{D})$. For a predictor f, let $\text{level}(f)$ be the partition of the domain $\mathcal{X}$ given by its level sets. By the definition of calibration, $f \in \text{cal}(\mathcal{D})$ iff it is equal to $g_{\text{level}(f)}$, which establishes the injection.

4.1 Desiderata for Calibration Measures

A calibration measure μ is a function that for a given distribution $\mathcal{D}$, maps predictors f in $\mathcal{F}_\mathcal{X}$ to values in $[0,1]$. We denote this value as $\mu_\mathcal{D}(f)$. At the bare minimum, we want $\mu_\mathcal{D}$ to satisfy completeness and soundness, meaning that for all $\mathcal{D}, f$,

$$\mu_\mathcal{D}(f) = 0 \text{ if } f \in \text{cal}(\mathcal{D}) \qquad \text{(Completeness)}$$
$$\mu_\mathcal{D}(f) > 0 \text{ if } f \notin \text{cal}(\mathcal{D}) \qquad \text{(Soundness)}$$

Ideally, we want these guarantees to be robust: $\mu(f)$ is small if f is close to calibrated, and large is f is far from calibrated. Formalizing this requires us to specify how we wish to measure the distance from calibration. A family of metrics m is a collection of metrics

[1] [14] introduced a notion of "smooth calibration" with an unrelated definition, but thankfully, they proved that their "smooth calibration" is in fact polynomially related to the [19] notion — therefore in our framework it is also a consistent calibration measure.

[2] We will assume that the domain $\mathcal{X}$ is discrete but possibly very large. As a consequence, $\text{Im}(f)$ is discrete, and events such as $f(x) = v$ for $v \in \text{Im}(f)$ are well defined. We can think of the finiteness assumption reflecting the fact that inputs to any model have finite precision. We do this to avoid measure-theoretic intricacies, but assuming $f : \mathcal{X} \to [0,1]$ is measurable should suffice when $\mathcal{X}$ is infinite.

[3] Consider the uniform distribution on $\mathcal{X} = \{0,1\}$ and let $f^*(x) = 1/2$ so labels are drawn uniformly. Consider the predictors $f(x) = x$ and $g(x) = 1 - x$. While $\ell_1(f,g) = 1$, the distributions $\mathcal{D}_f$ and $\mathcal{D}_g$ are identical, since f/g is uniform on $\{0,1\}$, and the labels are uniform conditioned on f/g.

$m_{\mathcal{D}}$ on $\mathcal{F}_{\mathcal{X}}$ for every distribution $\mathcal{D}$ on $\mathcal{X}$. For instance, the ℓ_p distance on $\mathcal{F}_{\mathcal{X}}$ under distribution $\mathcal{D}$ for $p \geq 1$ is given by

$$\ell_{p,\mathcal{D}}(f,g) = \mathop{\mathbb{E}}_{\mathcal{D}}[|f(x) - g(x)|^p]^{1/p}$$

We note that $m_{\mathcal{D}}$ only ought to depend on the marginal $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$. When the distribution $\mathcal{D}$ is clear, we will sometimes suppress the dependence on the distribution and refer to m as a metric rather than a family. Indeed, it is common to refer to the above distance as ℓ_p distance, ignoring the dependence on $\mathcal{D}$.

Definition 4.1 (True distance to calibration). Given a metric family m on $\mathcal{F}_{\mathcal{X}}$, we define the true m-distance to calibration under $\mathcal{D}$ as

$$\mathsf{dCE}^m_{\mathcal{D}}(f) = \min_{g \in \mathsf{cal}(\mathcal{D})} m_{\mathcal{D}}(f,g).$$

With this definition in place, we define consistent calibration measures with respect to m.

Definition 4.2 (Consistent calibration measures). For $c, s \geq 0$, we say that μ satisfies c-robust completeness w.r.t. m if there exist a constant $a \geq 0$ such that for every distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$, and predictor $f \in \mathcal{F}_{\mathcal{X}}$

$$\mu_{\mathcal{D}}(f) \leq a(\mathsf{dCE}^m_{\mathcal{D}}(f))^c \qquad \text{(Robust completeness)}$$

and s-robust soundness w.r.t. m if there exist $b \geq 0$ such that for every distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$, and predictor $f \in \mathcal{F}_{\mathcal{X}}$

$$\mu_{\mathcal{D}}(f) \geq b(\mathsf{dCE}^m_{\mathcal{D}}(f))^s. \qquad \text{(Robust soundness)}$$

We say that μ is an (c,s)-consistent calibration measure w.r.t m if both these conditions hold, and we define its approximation degree to be s/c. [4] We say μ it is an consistent calibration measure w.r.t. m if there exists $c, s \geq 0$ for which (c,s)-consistency for m holds.

To see that these names indeed make sense, observe that if f is ε-close to being perfectly calibrated, then robust completeness ensures that $\mu_{\mathcal{D}}(f)$ is $O(\varepsilon^c)$ and hence goes to 0 with ε. Robust soundness ensures that if $\mu_{\mathcal{D}}(f) = \varepsilon \to 0$, then $\mathsf{dCE}^m_{\mathcal{D}}(f) = O(\varepsilon^{1/s}) \to 0$. Conversely, when the true m-distance to calibration for f is $\eta \gg 0$, robust soundness ensures that $\mu_{\mathcal{D}}(f) = \Omega(\eta^s)$ is also bounded away from 0.

Given a sequence of predictors $\{f_n\}$, we say that the sequence converges to $f \in \mathcal{F}_{\mathcal{X}}$, denote $F_n \to f$ if

$$\lim_{n \to \infty} m_{\mathcal{D}}(f_n, f) = 0.$$

Robust soundness ensures that if $f_n \to g \in \mathsf{cal}(\mathcal{D})$, then $\mu(f_n) \to 0$. $\mathsf{dCE}^m_{\mathcal{D}}$ satisfies a stronger continuity property, namely that it is 1-Lipshcitz with respect to $m_{\mathcal{D}}$:

$$|\mathsf{dCE}^m_{\mathcal{D}}(f) - \mathsf{dCE}^m_{\mathcal{D}}(f')| \leq m_{\mathcal{D}}(f,f').$$

This property is easy to verify from the definition. It implies that for any $f \in \mathcal{F}_{\mathcal{X}}$ not necessarily calibrated, if $f_n \to f$, $\mu(f_n) \to \mu(f)$. Not every ℓ_1-consistent calibration measure have this stronger property of convergence everywhere, although some do.

Indeed, the following lemma implies that among all calibration measures that satisfy completeness and are 1-Lipschitz with respect to $m_{\mathcal{D}}$, $\mathsf{dCE}^m_{\mathcal{D}}$ is the largest. Thus any consistent calibration measure that can grow as $\omega(\mathsf{dCE}^m)$ cannot be Lipschitz.

[4] For metrics that can take on arbitrarily small values (such as the ℓ_p metrics), it follows that $s > c$.

Lemma 4.3. *Any calibration measure $\mu_{\mathcal{D}}$ which satisfies completeness and is L-Lipschitz w.r.t $m_{\mathcal{D}}$ must satisfy $\mu_{\mathcal{D}}(f) \leq L\ \mathsf{dCE}_{\mathcal{D}}(f)$ for all $f \in \mathcal{F}_{\mathcal{X}}$.*

Metric families that are particularly important to us are the ℓ_p metrics. Since all ℓ_p measures are polynomially related, the set of ℓ_p consistent calibration metrics is independent of p for bounded p.

Lemma 4.4. *The set of ℓ_p-consistent calibration measures is identical for all $p \in [1, \infty)$.*

Given this result, we will focus on the ℓ_1 metric and define the true distance to calibration by

$$\mathsf{dCE}_{\mathcal{D}}(f) := \mathsf{dCE}^{\ell_1}_{\mathcal{D}}(f) = \min_{g \in \mathsf{cal}(\mathcal{D})} \ell_{1,\mathcal{D}}(f,g).$$

(True distance from calibration)

It has good continuity properties, and the resulting set of consistent calibration measures does not depend on the choice of the ℓ_p metric. Henceforth when we refer to (c,s)-consistent calibration metrics without making m explicit, it is assumed that we mean the ℓ_1 distance. We note that there might be settings (not considered in this work) where other metrics on $\mathcal{F}_{\mathcal{X}}$ are suitable.

4.2 Approximation Limits in the PA Model

Given the desirable properties of dCE, one might wonder: why not use dCE as a calibration measure in itself? The main barrier to this is that dCE cannot be computed (or even defined) in the prediction access model. Indeed, if it were, there would be no need to look for alternative notions of approximate calibration.

Lemma 4.5. *Let $\alpha \in (0, 1/2]$. There exists a domain $\mathcal{X}$, a predictor $f \in \mathcal{F}_{\mathcal{X}}$, and distributions $\mathcal{D}^1, \mathcal{D}^2$ on $\mathcal{X} \times \{0,1\}$ such that*

- *The distributions $\mathcal{D}^1_f$ and $\mathcal{D}^2_f$ are identical.*
- *$\mathsf{dCE}_{\mathcal{D}^1}(f) \leq 2\alpha^2$, while $\mathsf{dCE}_{\mathcal{D}^2}(f) \geq \alpha$.*

This leads us to the quest for approximations that can be computed (efficiently) in the Prediction-access model. It implies that one can at best hope to get a degree 2 approximation to dCE.

Corollary 4.6. *Let $\mu(f)$ be a (ℓ_1, c, s)-consistent calibration measure computable in the prediction access model. Then $s \geq 2c$.*

Given this setup, we can now state our desiderata for an ideal calibration measure μ.

(1) **Access:** $\mu_{\mathcal{D}}(f) = \mu(\mathcal{D}_f)$ is well defined in the Prediction-only access model.
(2) **Consistency:** It is (c,s)-consistent. Ideally, it has degree $s/c = 2$.
(3) **Efficiency:** It can be computed within accuracy ε in time $\mathrm{poly}(1/\varepsilon)$ using $\mathrm{poly}(1/\varepsilon)$ random samples from $\mathcal{D}_f$.

4.3 On ECE and Other Measures

Recall that for a predictor f, we define its expected calibration error $\mathsf{ECE}(f)$ as

$$\mathsf{ECE}(f) = \mathbb{E}[|\mathbb{E}[y|f] - f|].$$

Clearly, ECE is well defined in the PA model. We analyze ECE in our framework and show that it satisfies 1-robust soundness, but not robust completeness. For the former, we present an alternate view of ECE in terms of ℓ_1 distance. Recall that $\mathrm{level}(f)$ is the partition

of $\mathcal{X}$ into the level sets of f, and that for a partition $\mathcal{S} = \{S_i\}$, the predictor $g_{\mathcal{S}}$ maps each $x \in S_i$ to $\mathbb{E}[y|S_i]$.

Lemma 4.7. *Let $\mathcal{S} = \text{level}(f)$. We have $\mathsf{ECE}(f) = \ell_1(f, g_{\mathcal{S}}) \geq \mathsf{dCE}_{\mathcal{D}}(f)$.*

The main drawbacks of ECE are that it does not satisfy robust completeness, and is discontinuous at 0.

Lemma 4.8. *$\mathsf{ECE}_{\mathcal{D}}(f)$ does not satisfy c-robust completeness for any $c > 0$. It can be discontinuous at 0.*

Table 1 summarize how other calibration measures that have been studied in the literature fare under our desiderata. Further discussion of these measures can be found in Appendix A of the full version [5].

5 DISTANCE BASED MEASURES IN THE PA MODEL

We start by defining upper and lower bounds to the true distance to calibration in the PA model. Our main result in this subsection is Theorem 5.5 showing that these are the *best possible bounds* one can have on dCE in the PA model. To define and analyze these distances, we need some auxiliary notions.

Definition 5.1. Let Γ be a distribution over $[0,1] \times \{0,1\}$. Define the set $\text{ext}(\Gamma)$ to consist of all joint distributions Π of triples $(u, v, y) \in [0,1] \times [0,1] \times \{0,1\}$, such that

- the marginal distribution of (v, y) is Γ;
- the marginal distribution (u, y) is perfectly calibrated: $\mathbb{E}_{\Pi}[y|u] = u$.

We define $\text{lift}(\Gamma)$ to be all pairs $(\mathcal{D}, f)$ where

- $\mathcal{D}$ is a distribution over $\mathcal{X} \times \{0,1\}$ for some domain $\mathcal{X}$.
- $f : \mathcal{X} \to [0,1]$ is predictor so that $\mathcal{D}_f = \Gamma$.

We first define the upper distance to calibration.

Definition 5.2 (Upper distance to calibration). For a distribution Γ over $[0,1] \times \{0,1\}$,let $K(\Gamma)$ denote the set of transformations $\kappa : [0,1] \to [0,1]$ such that the distribution of $(\kappa(v), y)$ for $(v, y) \sim \Gamma$ is perfectly calibrated. We define the *upper distance from calibration* $\overline{\mathsf{dCE}}(\Gamma)$ as

$$\overline{\mathsf{dCE}}(\Gamma) = \inf_{\kappa \in K(\Gamma)} \mathop{\mathbb{E}}_{(v,y)\sim\Gamma} [|v - \kappa(v)|],$$

For a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a predictor $f : \mathcal{X} \to [0,1]$, we define the *upper distance from calibration* $\overline{\mathsf{dCE}}_{\mathcal{D}}(f)$ to be $\overline{\mathsf{dCE}}(\mathcal{D}_f)$, or equivalently,

$$\overline{\mathsf{dCE}}_{\mathcal{D}}(f) := \inf_{\substack{\kappa:[0,1]\to[0,1] \\ \kappa\circ f \in \text{cal}(\mathcal{D})}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [|f(x) - \kappa(f(x))|].$$

We call this the upper distance since we only compare f with a calibrated predictor $\kappa \circ f$ that can be obtained by applying a postprocessing κ to f. It follows immediately that $\overline{\mathsf{dCE}}_{\mathcal{D}}(f) \geq \mathsf{dCE}_{\mathcal{D}}(f)$.

We next define the lower distance to calibration.

Definition 5.3 (Lower distance to calibration). We define the *lower distance to calibration* denoted $\underline{\mathsf{dCE}}(\Gamma)$ as

$$\underline{\mathsf{dCE}}(\Gamma) := \inf_{\Pi \in \text{ext}(\Gamma)} \mathop{\mathbb{E}}_{(u,v,y)\sim\Pi} |u - v|. \tag{7}$$

For a distribution $\mathcal{D}$ and a predictor f, we define $\underline{\mathsf{dCE}}_{\mathcal{D}}(f) := \underline{\mathsf{dCE}}(\mathcal{D}_f)$.

The following lemma justifies the terminology of upper and lower distance.

Lemma 5.4. *We have $\underline{\mathsf{dCE}}_{\mathcal{D}}(f) \leq \mathsf{dCE}_{\mathcal{D}}(f) \leq \overline{\mathsf{dCE}}_{\mathcal{D}}(f)$*

Proof. Every calibrated predictor $g \in \text{cal}(\mathcal{D})$ gives a distribution $\Pi \in \text{ext}(\mathcal{D}_f)$ where we sample $(x, y) \sim \mathcal{D}$ and return $(g(x), f(x), y)$. Note that $\mathbb{E}_{(u,v,y)\sim\Pi}[|u - v|] = \ell_1(f, g)$. Minimizing over $g \in \text{cal}(\mathcal{D})$ gives the first inequality. The second follows because in the definition of $\overline{\mathsf{dCE}}$ we minimize over a subset of $\text{cal}(\mathcal{D})$, namely only those $g = \kappa \circ f$ that can be obtained from f via a postprocessing κ. □

We now show that these are the best possible bounds one can have on dCE in the PA model.

Theorem 5.5. *The following identities hold*

$$\underline{\mathsf{dCE}}(\Gamma) = \inf_{(\mathcal{D},f)\in\text{lift}(\Gamma)} \mathsf{dCE}_{\mathcal{D}}(f)$$

$$\overline{\mathsf{dCE}}(\Gamma) = \sup_{(\mathcal{D},f)\in\text{lift}(\Gamma)} \mathsf{dCE}_{\mathcal{D}}(f).$$

The gap between each of these quantities can be at least quadratic, as the distributions $\mathcal{D}^1$ and $\mathcal{D}^2$ in Lemma 4.5 shows. Under both distributions $\mathcal{D}^1$ and $\mathcal{D}^2$, we have $\underline{\mathsf{dCE}}(f) \leq 2\alpha^2, \overline{\mathsf{dCE}}(f) = \alpha$. But $\mathsf{dCE}_{\mathcal{D}^1}(f) = 2\alpha^2$ while $\mathsf{dCE}_{\mathcal{D}^2}(f) = \alpha$. We will show that this gap is indeed tight in the next section using the notion of Interval calibration.

6 INTERVAL CALIBRATION

In this section, we introduce the notion of interval calibration. Our main result is Theorem 6.2 which shows that it is quadratically related to the true distance from calibration. Since intCE is defined in the PA model, this implies in particular, that there might be at most quadratic gap between $\underline{\mathsf{dCE}}$ and $\overline{\mathsf{dCE}}$. We exhibit a gap instance showing that this is tight (Lemma 6.5). As defined, it is unclear if Interval calibration can be efficiently estimated. We propose a surrogate version of interval calibration which gives similar bounds and can be efficiently estimated from samples in Section 6.2 of the full version [5].

An interval partition $\mathcal{I}$ of $[0,1]$ is a partition of $[0,1]$ into disjoint intervals $\{I_j\}_{j\in[m]}$. Let $w(I)$ denote the width of interval I.

Definition 6.1 (Interval Calibration Error). For a distribution Γ over $[0,1] \times \{0,1\}$ and interval partition $\mathcal{I}$ define

$$\mathsf{binnedECE}(\Gamma, \mathcal{I}) := \sum_{j\in[m]} \left| \mathop{\mathbb{E}}_{(v,y)\sim\Gamma} [(v - y)\mathbf{1}(v \in I_j)] \right|.$$

We define the *average interval width*

$$w_{\Gamma}(\mathcal{I}) := \sum_{j\in[m]} \mathop{\mathbb{E}}_{(v,y)\sim\Gamma} [\mathbf{1}(v \in I_j) w(I_j)].$$

The *interval calibration error* $\mathsf{intCE}(\Gamma)$ is then the minimum of $\mathsf{binnedECE}(\Gamma, \mathcal{I}) + w_{\Gamma}(\mathcal{I})$ over all interval partitions $\mathcal{I}$:

$$\mathsf{intCE}(\Gamma) := \min_{\mathcal{I}:\text{ Interval partition}} \left(\mathsf{binnedECE}(\Gamma, \mathcal{I}) + w_{\Gamma}(\mathcal{I})\right).$$

For a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a predictor $f : \mathcal{X} \to [0,1]$, we define $\mathsf{intCE}_{\mathcal{D}}(f, \mathcal{I}) := \mathsf{intCE}(\mathcal{D}_f, \mathcal{I})$ and $\mathsf{intCE}_{\mathcal{D}}(f) := \mathsf{intCE}(\mathcal{D}_f)$.

Our main theorem about interval calibration is the following.

Theorem 6.2. *We have* $\overline{\mathsf{dCE}}(\Gamma) \le \mathsf{intCE}(\Gamma) \le 4\sqrt{\underline{\mathsf{dCE}}(\Gamma)}$.

Combining this with Lemma 5.4, we conclude that intCE is indeed a quadratic approximation to the true distance from calibration, which is the best achievable in the PA model by Corollary 4.6.

Corollary 6.3. intCE *is a* $(1/2, 1)$*-consistent calibration measure. We have*

$$\mathsf{dCE}_{\mathcal{D}}(f) \le \mathsf{intCE}_{\mathcal{D}}(f) \le 4\sqrt{\mathsf{dCE}_{\mathcal{D}}(f)}.$$

Another corollary is the following bounds for distance measures, which shows that the gaps presented in Lemma 4.5 are the largest possible.

Corollary 6.4. *We have*

$$\underline{\mathsf{dCE}}_{\mathcal{D}}(f) \le \mathsf{dCE}_{\mathcal{D}}(f) \le 4\sqrt{\underline{\mathsf{dCE}}_{\mathcal{D}}(f)},$$
$$\frac{1}{16}\overline{\mathsf{dCE}}_{\mathcal{D}}(f)^2 \le \mathsf{dCE}_{\mathcal{D}}(f) \le \overline{\mathsf{dCE}}_{\mathcal{D}}(f).$$

Quadratic Gap between Interval Calibration and Upper Calibration Distance. For a distribution $\mathcal{D}$ and a predictor f, our results in previous subsections imply the following chain of inequalities (omitting $\mathcal{D}$ in the subscript for brevity):

$$\underline{\mathsf{dCE}}(f) \le \overline{\mathsf{dCE}}(f) \le \mathsf{intCE}(f) \le 4\sqrt{\underline{\mathsf{dCE}}(f)}. \tag{8}$$

These inequalities completely characterize the relationship between $\underline{\mathsf{dCE}}(f)$ and $\overline{\mathsf{dCE}}(f)$ and also the relationship between $\underline{\mathsf{dCE}}(f)$ and $\mathsf{intCE}(f)$ for the following reason. By Lemma 4.5, we know that $\overline{\mathsf{dCE}}(f)$ can be as large as $\Omega(\sqrt{\underline{\mathsf{dCE}}(f)})$, which implies that $\mathsf{intCE}(f)$ can be as large as $\Omega(\sqrt{\underline{\mathsf{dCE}}(f)})$. Also, it is easy to show that $\mathsf{intCE}(f)$ can be as small as $O(\underline{\mathsf{dCE}}(f))$ by choosing f to be a constant function, which implies that $\overline{\mathsf{dCE}}(f)$ can be as small as $O(\underline{\mathsf{dCE}}(f))$.

The remaining question is whether (8) completely characterizes the relationship between $\overline{\mathsf{dCE}}(f)$ and $\mathsf{intCE}(f)$. We show that the answer is yes by the following lemma (Lemma 6.5) which gives examples where $\mathsf{intCE}(f) = \Omega((\overline{\mathsf{dCE}}(f))^{1/2})$. We also show that $\mathsf{intCE}(f)$ can be discontinuous as a function of f in Lemma 6.6.

Lemma 6.5. *For any* $\alpha \in (0, 1/4)$*, there exist distribution* $\mathcal{D}$ *and predictor* f *such that*

$$\overline{\mathsf{dCE}}_{\mathcal{D}}(f) \le 5\alpha^2 \quad \text{and} \quad \mathsf{intCE}_{\mathcal{D}}(f) \ge \alpha.$$

Lemma 6.6. *There exist a distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \{0,1\}$ *and a sequence of predictors* $f_n : \mathcal{X} \to [0,1]$ *converging uniformly to a predictor* $f : \mathcal{X} \to [0,1]$ *as* $n \to \infty$ *such that* $\lim_{n\to\infty} \mathsf{intCE}_{\mathcal{D}}(f_n) \neq \mathsf{intCE}_{\mathcal{D}}(f)$.

7 SMOOTH CALIBRATION AND THE LOWER DISTANCE TO CALIBRATION

In this section, we define and analyze the notion of smooth calibration. The main result of this section is that the smooth calibration error smCE is equivalent, up to a constant factor, to $\underline{\mathsf{dCE}}$. We also give algorithms that can compute both these quantities to within an additive ε in time $\mathrm{poly}(1/\varepsilon)$ on an empirical distribution.

At a high level, the proof that $\underline{\mathsf{dCE}}$ and smCE are related proceeds as follows.

(1) For a distribution Γ over $[0,1] \times \{0,1\}$, our definition of $\underline{\mathsf{dCE}}(\Gamma)$ (Definition 5.3) is based on couplings $\Pi \in \mathrm{ext}(\Gamma)$ that connect Γ to calibrated distributions Γ'. For a given distribution $\mathcal{D}$, the space of predictors $f : \mathcal{X} \to [0,1]$ which are calibrated is non-convex (for finite $\mathcal{X}$, it is a finite set). But when we move to the space of distributions Γ' over $[0,1] \times \{0,1\}$, then the space of perfectly calibrated distributions is convex. This is because for $(v, b) \in [0,1] \times \{0,1\}$ if $\Gamma'(v, b)$ denotes the probability assigned to it, then the calibration constraint states that for every v,
$$\frac{\Gamma'(v,1)}{\Gamma'(v,0) + \Gamma'(v,1)} = v$$
which is a linear constraint for every v. This allows us to view the problem of computing $\underline{\mathsf{dCE}}$ as optimization over couplings Π connecting Γ to some Γ' satisfying such linear constraints.
(2) We show that by suitably discretizing $[0,1]$, we can write the problem of computing $\underline{\mathsf{dCE}}$ as a linear program. The dual of this program (after some manipulation) asks for a 2-Lipschitz function $w : [0,1] \to [-1,1]$ which witnesses the lack of calibration of f, by showing that $\mathbb{E}[w(v)(y - v)]$ is large. Rescaling gives a 1-Lipschitz function which proves that $\mathsf{smCE}_{\mathcal{D}}(f) \ge \underline{\mathsf{dCE}}_{\mathcal{D}}(f)/2$. The other direction which corresponds to weak duality is easy to show.

We now proceed with the formal definitions and proof. We start by defining a general family of calibration measures called *weighted calibration error* from [19].

Definition 7.1 (Weighted calibration). [19] Let W be a family of functions $w : [0,1] \to \mathbb{R}$. The *weighted calibration error* of a distribution Γ over $[0,1] \times \{0,1\}$ is defined as

$$\mathsf{wCE}^{W}(\Gamma) := \sup_{w \in W} \left| \mathop{\mathbb{E}}_{(v,y)\sim\Gamma} [(y - v)w(v)] \right|.$$

Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and predictor $f : \mathcal{X} \to [0,1]$, we denote the weighted calibration error of f under $\mathcal{D}$ as

$$\mathsf{wCE}^{W}_{\mathcal{D}}(f) := \mathsf{wCE}^{W}(\mathcal{D}_f) = \sup_{w \in W} \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [(y - f(x))w(f(x))] \right|.$$

Clearly, any weighted calibration error notion is well defined in the PA model. Moreover, all of those at the very least satisfy completeness: if Γ is perfectly calibrated, than for any w, we have

$$\mathop{\mathbb{E}}_{\Gamma}[(y - v)w(v)] = \mathbb{E}[\mathbb{E}[(y - v)w(v)|v]],$$

and since $\mathbb{E}[y|v] = v$ for a perfectly calibrated predictor, this latter quantity is zero.

A particularly important calibration measure among those is the *smooth calibration* where W is the family of all 1-Lipschitz, bounded functions. This was introduced in the work of [25] who termed it *weak calibration*, the terminology of smooth calibration is from [19].

Definition 7.2 (Smooth calibration). Let L be the family of all 1-Lipschitz functions $w : [0,1] \to [-1,1]$. The *smooth calibration error* of a distribution Γ over $[0,1] \times \{0,1\}$ is defined as weighted calibration error with respect to the family of all 1-Lipschitz functions

$$\mathsf{smCE}(\Gamma) := \mathsf{wCE}^{L}(\Gamma). \tag{9}$$

Accordingly, for a distribution $\mathcal{D}$ and a predictor f, we define

$$\mathsf{smCE}_{\mathcal{D}}(f) := \mathsf{smCE}(\mathcal{D}_f) = \mathsf{wCE}^{L}(\mathcal{D}_f) = \mathsf{wCE}^{L}_{\mathcal{D}}(f).$$

Our main result on the smooth calibration error is the following.

Theorem 7.3. *For any distribution Γ over $[0,1] \times \{0,1\}$, we have*

$$\frac{1}{2}\underline{\mathsf{dCE}}(\Gamma) \le \mathsf{smCE}(\Gamma) \le 2\underline{\mathsf{dCE}}(\Gamma).$$

Combining this with corollary 6.4, we conclude that smCE is a $(1,2)$-consitent calibration measure, and yields an optimal degree-2 approximation to dCE. Along the way, we will find an efficient algorithm for computing $\underline{\mathsf{dCE}}(\Gamma)$ (see Remark 7.10).

As it is often the case, the inequality $\mathsf{smCE}(\Gamma) \le 2\underline{\mathsf{dCE}}(\Gamma)$ is significantly easier to prove (corresponding to the weak duality). We will start by proving this easier direction. The following lemma is a strengthening of the fact that smCE is Lipschitz continuous in the predictor f (i.e. for two predictors f, g defined on the same set $\mathcal{X}$, we have $|\mathsf{smCE}_{\mathcal{D}}(f) - \mathsf{smCE}_{\mathcal{D}}(g)| \le 2\ell_1(f,g)$).

Lemma 7.4. *Let Π be a distribution over $[0,1] \times [0,1] \times \{0,1\}$. For any 1-Lipschitz function $w : [0,1] \to [-1,1]$, we have*

$$\left| \mathop{\mathbb{E}}_{(u,v,y)\sim\Pi}[(y-u)w(u)] - \mathop{\mathbb{E}}_{(u,v,y)\sim\Pi}[(y-v)w(v)] \right| \le 2 \mathop{\mathbb{E}}_{(u,v,y)\sim\Pi}|u-v|.$$

In the full version [5], we use the lemma above to prove the upper bound on $\mathsf{smCE}(\Gamma)$ in Theorem 7.3. To prove Theorem 7.3, it remains to prove the lower bound on $\mathsf{smCE}(\Gamma)$. We prove that in the rest of the section.

7.1 Linear Program Formulation of Lower Calibration Distance

For a distribution Γ over $[0,1] \times \{0,1\}$, we show that a discretized version of $\underline{\mathsf{dCE}}(\Gamma)$ can be formulated as the optimal value of a linear program, and the error caused by the discretization can be made arbitrarily small. We then use the strong duality theorem of linear programming to prove the lower bound of $\mathsf{smCE}(\Gamma)$ in Theorem 7.3. The linear programming formulation also allows us to give an alternative proof of the upper bound in Theorem 7.3 using the weak duality theorem. Moreover, the linear program formulation gives us an efficient algorithm for estimating $\underline{\mathsf{dCE}}(\Gamma)$.

Our first step is to assume that Γ is a distribution over $V \times \{0,1\}$ for some *finite* set $V \subseteq [0,1]$. This is mostly without loss of generality because for $\varepsilon > 0$ we can round every value $v \in [0,1]$ in $(v,y) \sim \Gamma$ to the closest value in $\{0, \varepsilon, 2\varepsilon, \ldots\} \cap [0,1]$ without changing $\underline{\mathsf{dCE}}(\Gamma)$ by more than ε. The following definition allows us to further define discretized versions of $\mathsf{ext}(\Gamma)$ and $\underline{\mathsf{dCE}}(\Gamma)$:

Definition 7.5. Let $U, V \subseteq [0,1]$ be finite sets. Let Γ be a distribution over $V \times \{0,1\}$. Define the set $\mathsf{ext}^{U}(\Gamma)$ to consist of all joint distributions Π of triples $(u,v,y) \in U \times V \times \{0,1\}$, such that

- the marginal distribution of (v,y) is Γ;
- the marginal distribution (u,y) is perfectly calibrated: $\mathbb{E}_{\Pi}[y|u] = u$.

We define $\underline{\mathsf{dCE}}^{U}(\Gamma)$ to be

$$\underline{\mathsf{dCE}}^{U}(\Gamma) := \inf_{\Pi \in \mathsf{ext}^{U}(\Gamma)} \mathop{\mathbb{E}}_{(u,v,y)\sim\Pi} |u-v|. \tag{10}$$

Later in Lemma 7.11 we will show that $\underline{\mathsf{dCE}}^{U}(\Gamma)$ is close to $\underline{\mathsf{dCE}}(\Gamma)$ as long as U is a suitably rich class. For now, we show how to formulate $\underline{\mathsf{dCE}}^{U}(\Gamma)$ as the optimal value of a linear program:

Lemma 7.6. *Let U, V, Γ be defined as in Definition 7.5 and assume $\{0,1\} \subseteq U$. By a slight abuse of notation, we define $\Gamma(v,y)$ to be the probability mass of Γ on $(v,y) \in V \times \{0,1\}$. Then the following linear program with variables $\Pi(u,v,y)$ for $(u,v,y) \in U \times V \times \{0,1\}$ is feasible and its optimal value equals $\underline{\mathsf{dCE}}^{U}(\Gamma)$:*

$$\begin{aligned}
\text{minimize} \quad & \sum_{(u,v,y)\in U\times V\times\{0,1\}} |u-v|\,\Pi(u,v,y) && (11)\\
\text{s.t.} \quad & \sum_{u\in U} \Pi(u,v,y) = \Gamma(v,y) \\
& \qquad \text{for every } (v,y) \in V \times \{0,1\}; && (r(v,y))\\
& (1-u)\sum_{v\in V} \Pi(u,v,1) = u \sum_{v\in V} \Pi(u,v,0) \\
& \qquad \text{for every } u \in U; && (s(u))\\
& \Pi(u,v,y) \ge 0 \\
& \qquad \text{for every } (u,v,y) \in U \times V \times \{0,1\}.
\end{aligned}$$

Moreover, the dual of the linear program (11) is the following linear program (12) with variables $r(v,y)$ and $s(u)$ for $u \in U, v \in V$ and $y \in \{0,1\}$. By the duality theorem, the optimal value of (12) is also $\underline{\mathsf{dCE}}^{U}(\Gamma)$.

$$\begin{aligned}
\text{maximize} \quad & \sum_{(v,y)\in V\times\{0,1\}} r(v,y)\Gamma(v,y) && (12)\\
\text{s.t.} \quad & r(v,y) \le |u-v| + (y-u)s(u) \\
& \qquad \text{for every } (u,v,y) \in U \times V \times \{0,1\}. && (\Pi(u,v,y))
\end{aligned}$$

Proof. Any distribution Π over $U \times V \times \{0,1\}$ corresponds to a function $\Pi : U \times V \times \{0,1\} \to \mathbb{R}$ where $\Pi(u,v,y)$ is the probability mass on $(u,v,y) \in U \times V \times \{0,1\}$. It is easy to check that if the distribution Π belongs to $\mathsf{ext}(\Gamma)$, then the function Π satisfies the constraints of (11), and conversely, any function Π satisfying the constraints also corresponds to a distribution $\Pi \in \mathsf{ext}(\Gamma)$. In particular, for $(u,v,y) \sim \Pi$, the first constraint ensures that the marginal distribution of (v,y) is Γ, and the second constraint ensures that the marginal distribution of (u,y) is calibrated. Moreover, the objective of (11) corresponds to the expectation $\mathbb{E}_{(u,v,y)\sim\Pi}|u-v|$ in (10). This proves that $\underline{\mathsf{dCE}}^{U}(\Gamma)$ is equal to the optimal value of the linear program (11). To show that the linear program (11) is feasible, consider setting $\Pi(u,v,y) = \Gamma(v,y)$ if $u = y$, and setting $\Pi(u,v,y) = 0$ if $u \ne y$. It is easy to check that this choice of Π satisfies the constraints of (11) using our assumption that $\{0,1\} \subseteq U$. □

Claim 7.7. *Let $U, V \subseteq [0,1]$ be finite sets and assume $\{0,1\} \subseteq U$. The optimal value of the dual linear program (12) does not change*

even if we add the additional constraints $-1 \le s(u) \le 1$ *for every* $u \in U$.

Remark 7.8. Since $\Gamma(v,y)$ in the objective of (12) is nonnegative, it is always without loss of generality to assume that $r(v,y)$ is as large as possible, i.e.,

$$r(v,y) = \min_{u \in U} \big(|u-v| + (y-u)s(u)\big). \tag{13}$$

Assuming (13), it is easy to check that $r(v,y)$ is 1-Lipschitz in v, i.e., $|r(v_1,y) - r(v_2,y)| \le |v_1 - v_2|$ for every $v_1, v_2 \in V$ and $y \in \{0,1\}$. When $\{0,1\} \subseteq U$, Claim 7.7 allows us to assume that $-1 \le s(u) \le 1$ without loss of generality. When this assumption and (13) are both satisfied, it is easy to verify that $r(v,y) \in [-|v-y|, |v-y|] \subseteq [-1,1]$ and $r(v,1) - r(v,0) \in [-1,1]$ for every $v \in V$ and $y \in \{0,1\}$. Indeed, in (13) we have $|u-v| + (y-u)s(u) \ge |u-v| - |y-u| \ge -|v-y|$ and thus $r(v,y) \ge -|v-y|$. The upper bound $r(v,y) \le |v-y|$ can be proved by setting $u = y$ in (13) using our assumption $\{0,1\} \subseteq U$.

Remark 7.9. When U, V are finite sets satisfying $\{0,1\} \subseteq U = V \subseteq [0,1]$, using Remark 7.8 one can verify that the dual linear program (12) has the same optimal value as the following linear program:

$$\begin{aligned} \text{maximize} \quad & \sum_{(v,y) \in V \times \{0,1\}} r(v,y)\Gamma(v,y) && (14)\\ \text{s.t.} \quad & |r(v_1,y) - r(v_2,y)| \le |v_1 - v_2| \\ & \qquad \text{for every } (v_1,v_2,y) \in V \times V \times \{0,1\}; && (15)\\ & r(v,y) \le (y-v)s(v) \\ & \qquad \text{for every } (v,y) \in V \times \{0,1\}. \end{aligned}$$

The constraints (15) can be enforced simply by checking neighboring pairs (v_1, v_2) when the values in V are sorted. Thus the effective number of constraints in (15) is $O(|V|)$.

Remark 7.10. Let $U \subseteq [0,1]$ be a finite set satisfying $\{0,1\} \subseteq U$. Given a distribution Γ over $V \times \{0,1\}$ for a finite $V \subseteq [0,1]$, Lemma 7.6 allows us to efficiently compute $\underline{\mathsf{dCE}}^U(\Gamma)$ by solving either the primal linear program (11) or the dual linear program (12). When $U = V$, it may be more efficient to solve the equivalent linear program (14) which effectively has only $O(|V|)$ constraints as we mention in Remark 7.9. Moreover, given two distributions Γ and Γ' that are close in a certain Wasserstein distance, using the dual linear program (12) we can show that $\underline{\mathsf{dCE}}^U(\Gamma')$ and $\underline{\mathsf{dCE}}^U(\Gamma)$ are close (we make this formal in Lemma 9.11 of the full version [5]). This allows us to estimate $\underline{\mathsf{dCE}}^U(\Gamma)$ only using examples drawn from Γ (see Section 9.2 of the full version [5]). In Lemma 7.11 below we show that choosing $|U| = O(1/\varepsilon)$ suffices to ensure that $\underline{\mathsf{dCE}}^U(\Gamma)$ approximates $\underline{\mathsf{dCE}}(\Gamma)$ up to an additive error ε.

The following lemma relates $\underline{\mathsf{dCE}}^U(\Gamma)$ and $\underline{\mathsf{dCE}}(\Gamma)$:

Lemma 7.11. *Let* Γ *be a distribution over* $V \times \{0,1\}$ *for a finite* $V \subseteq [0,1]$. *Let* U *be a finite* ε *covering of* $[0,1]$ *satisfying* $\{0,1\} \subseteq U$. *That is, there exists* $\sigma : [0,1] \to U$ *such that* $|u - \sigma(u)| \le \varepsilon$ *for every* $u \in [0,1]$. *Then we have*

$$\underline{\mathsf{dCE}}(\Gamma) \le \underline{\mathsf{dCE}}^U(\Gamma) \le \underline{\mathsf{dCE}}(\Gamma) + 2\varepsilon.$$

We prove lower and upper bounds for $\mathsf{smCE}(\Gamma)$ using $\underline{\mathsf{dCE}}^U(\Gamma)$ in the two lemmas below.

Lemma 7.12. *Let* Γ *be a distribution over* $V \times \{0,1\}$ *for a finite* $V \subseteq [0,1]$. *Define* $U = V \cup \{0,1\}$. *Then* $\underline{\mathsf{dCE}}^U(\Gamma) \le 2\mathsf{smCE}(\Gamma)$.

Lemma 7.13. *Let* Γ *be a distribution over* $V \times \{0,1\}$ *for a finite* $V \subseteq [0,1]$. *For any finite* $U \subseteq [0,1]$, *we have* $\mathsf{smCE}(\Gamma) \le 2\underline{\mathsf{dCE}}^U(\Gamma)$.

In the proofs of Lemmas 7.12 and 7.13 above, we use the fact that $\underline{\mathsf{dCE}}^U(\Gamma)$ is equal to the optimal value of the dual linear program (12). However, for Lemma 7.12 we only need the fact that $\underline{\mathsf{dCE}}^U(\Gamma)$ is *at most* the optimal value, whereas for Lemma 7.13 we only need the fact that $\underline{\mathsf{dCE}}^U(\Gamma)$ is *at least* the optimal value. That is, our proof of Lemma 7.12 is based on the strong duality theorem, whereas the proof of Lemma 7.13 is based on the weak duality theorem. Below we apply Lemma 7.12 and Lemma 7.13 to prove the lower and upper bounds of $\mathsf{smCE}(\Gamma)$ in Theorem 7.3, respectively.

Proof of Theorem 7.3. For $\varepsilon_1 > 0$, we round the value $v \in [0,1]$ in $(v,y) \sim \Gamma$ to the closest value $v' \in \{0, \varepsilon_1, 2\varepsilon_1, \dots,\} \cap [0,1]$. Let Γ' be the distribution of (v', y). It is clear that $|\underline{\mathsf{dCE}}(\Gamma') - \underline{\mathsf{dCE}}(\Gamma)| \le \varepsilon_1$, and by Lemma 7.4 we have $|\mathsf{smCE}(\Gamma') - \mathsf{smCE}(\Gamma)| \le 2\varepsilon_1$.

By Lemma 7.11, for any $\varepsilon_2 > 0$, there exists a finite set $U \subseteq [0,1]$ such that $\underline{\mathsf{dCE}}(\Gamma') \le \underline{\mathsf{dCE}}^U(\Gamma') \le \underline{\mathsf{dCE}}(\Gamma') + \varepsilon_2$. Moreover, we can always choose U so that $\{0,1\} \cup V \subseteq U$. Now by Lemma 7.12,

$$\begin{aligned} \underline{\mathsf{dCE}}(\Gamma) - \varepsilon_1 \le \underline{\mathsf{dCE}}(\Gamma') \le \underline{\mathsf{dCE}}^U(\Gamma') &\le 2\mathsf{smCE}(\Gamma') \\ &\le 2\mathsf{smCE}(\Gamma) + 4\varepsilon_1. \end{aligned}$$

By Lemma 7.13,

$$\begin{aligned} \mathsf{smCE}(\Gamma) - 2\varepsilon_1 \le \mathsf{smCE}(\Gamma') &\le 2\underline{\mathsf{dCE}}^U(\Gamma') \\ &\le 2\underline{\mathsf{dCE}}(\Gamma') + 2\varepsilon_2 \\ &\le 2\underline{\mathsf{dCE}}(\Gamma) + 2\varepsilon_1 + 2\varepsilon_2. \end{aligned}$$

Taking $\varepsilon_1, \varepsilon_2 \to 0$ completes the proof. □

We conclude with an efficient algorithm for smooth calibration error. The generalization bound to accompany it will be proved in Corollary 9.9 in Section 9 of the full version [5].

Theorem 7.14. *For the empirical distribution* Γ *over a sample* $S = ((v_1,y_1), \dots (v_n,y_n)) \in ([0,1] \times \{0,1\})^n$ *we can calculate*

$$\mathsf{smCE}(\Gamma) := \sup_{w \in L} \frac{1}{n} \sum_i (y_i - v_i) w(v_i)$$

in time $\mathrm{poly}(n)$, *where* L *is the family of all* 1*-Lipschitz functions* $w : [0,1] \to [-1,1]$.

8 KERNEL CALIBRATION ERROR

We now consider kernel calibration (kCE^K), which is a special case of weighted calibration (Definition 7.1) where the family of weight functions lies in a Reproducing Kernel Hilbert Space $\mathcal{H}$. This notion was previously defined in [32] (called "MMCE"), motivated as a differentiable proxy for ECE.

We advance the theory of kernel calibration in several ways. First, we show that the kernel calibration error for the *Laplace* kernel is in fact a consistent calibration measure. This provides strong theoretical justification for measuring kernel calibration, and also gives a reason to use the *Laplace* kernel specifically, among other choices of kernel. Indeed, we complement the Laplace kernel

with a negative result: using the Gaussian kernel does not yield a consistent calibration measure.

Finally, as a curiosity, we observe that the techniques of [45] yield an alternate estimator for Laplace kernel calibration error, which bears similarity to the randomized-binning estimator of interval calibration error.

8.1 Preliminaries

We consider a *Reproducing Kernel Hilbert Space* of functions on a real line $\mathbb{R}$, i.e. a Hilbert space $\mathcal{H}$ of functions $h : \mathbb{R} \to \mathbb{R}$, with the associated norm $\|\cdot\|_{\mathcal{H}}$. This space is equipped with the feature map $\phi : \mathbb{R} \to \mathcal{H}$, satisfying $\langle h, \phi(v)\rangle_{\mathcal{H}} = h(v)$. The associated kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is now defined as $K(u, v) = \langle \phi(u), \phi(v)\rangle_{\mathcal{H}}$.

Definition 8.1 (Kernel Calibration Error [32]). Given a RKHS $\mathcal{H}$ with the norm $\|\cdot\|_{\mathcal{H}}$, we can consider a class of functions bounded by 1 with respect to this norm $B_{\mathcal{H}} := \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \le 1\}$, and we can study the associated weighted calibration error $\mathsf{wCE}^{B_{\mathcal{H}}}$ (as in Definition 7.1).

The *kernel calibration error* of a distribution Γ over $[0, 1] \times \{0, 1\}$ associated with the kernel K is defined as weighted calibration error with respect to the family of weight functions $B_{\mathcal{H}}$

$$\mathsf{kCE}^K(\Gamma) := \mathsf{wCE}^{B_{\mathcal{H}}}(\Gamma). \tag{16}$$

Accordingly, for a distribution $\mathcal{D}$ and a predictor f, we define $\mathsf{kCE}^{K,\mathcal{D}}(f) := \mathsf{kCE}_K(\mathcal{D}_f)$.

The following results are standard, from [32]. First, kCE_K can be written as the K-norm of a certain function, without explicitly maximizing over weight functions $h \in \mathcal{H}$.

Lemma 8.2 ([32]). *For any kernel K and the associated RKHS $\mathcal{H}$, and any distribution Γ over $[0, 1] \times \{0, 1\}$,*

$$\mathsf{kCE}^K(\Gamma) = \| \mathop{\mathbb{E}}_{(v,y)\sim\Gamma} [(y - v)\phi(v)] \|_{\mathcal{H}}.$$

This expression can be efficiently evaluated for an empirical distribution on a samples $S = \{(y_1, v_1), \dots (y_k, v_k)\}$.

Claim 8.3 ([32]). *Let Γ be the empirical distribution over a given sample $\{(v_1, y_1), \dots, (v_n, y_n)\}$. We can compute $\mathsf{kCE}^K(\Gamma)$ in time $O(n^2)$ using $O(n^2)$ evaluations of the kernel function:*

$$\mathsf{kCE}^K(\Gamma)^2 = \frac{1}{n^2} \sum_{i,j} (y_i - v_i)(y_j - v_j) K(v_i, v_j). \tag{17}$$

In Section 9 of the full version [5] we discuss the convergence of the kernel calibration error for the empirical distribution over the sample, to the kernel calibration error of the entire distribution — this convergence, together with Claim 8.3 gives an efficient way to estimate the kernel calibration error of a given predictor from a boudned number of samples from the underlying distribution.

The Laplace Kernel. We recall standard facts about the Laplace kernel $K_{\mathsf{Lap}}(u, v) := \exp(-|u - v|)$, and its associated RKHS. It turns out that the norm induced by functions in the associated RKHS has simple explicit expression — the corresponding space is a Sobolev space.

Fact 8.4 ([2]). *For the Laplace kernel $K_{\mathsf{Lap}}(u, v) = \exp(-|u - v|)$, we have associated RKHS $\mathcal{H}_{\mathsf{Lap}} = \{h : \mathbb{R} \to \mathbb{R} : \int \hat{h}(\omega)^2(1 + \omega^2)\,\mathrm{d}\omega < \infty\}$, where $\hat{h}$ denotes the Fourier transform of u. The associated inner product is given by*

$$\langle h_1, h_2\rangle_{K_{\mathsf{Lap}}} = \int_{-\infty}^{\infty} \hat{h}_1(\omega)\hat{h}_2(\omega)(1 + \omega^2)\,\mathrm{d}\omega,$$

in particular, for function $h : \mathbb{R} \to \mathbb{R}$,

$$\|h\|^2_{K_{\mathsf{Lap}}} = \int_{-\infty}^{\infty} \hat{h}(\omega)^2(1 + \omega^2)\,\mathrm{d}\omega = \|h\|_2^2 + \|h'\|_2^2.$$

8.2 Laplace Kernel Calibration Error Is a Consistent Calibration Measure

We now ask whether there is a kernel K for which kCE_K is a consistent calibration measure. The main result in this section is to show that this is the case for the *Laplace* kernel. Specifically, we prove that:

Theorem 8.5. *The Laplace kernel calibration error $\mathsf{kCE}^{\mathsf{Lap}} := \mathsf{kCE}^{K_{\mathsf{Lap}}}$ satisfies the following inequalities*

$$\frac{1}{3}\mathsf{smCE}(\Gamma) \le \mathsf{kCE}^{\mathsf{Lap}}(\Gamma) \le \sqrt{\underline{\mathsf{dCE}}(\Gamma)}.$$

By Corollary 6.4 and Theorem 7.3 it follows that $\mathsf{kCE}^{\mathsf{Lap}}$ is a $(1/2, 2)$-consistent calibration measure. Interestingly, the choice of kernel is crucial: we show that for the Gaussian kernel, the resulting measure does not satisfy robust soundness anymore. Specifically, we prove the following theorem.

Theorem 8.6. *For every ε, there is a distribution Γ_ε over $[0, 1] \times \{0, 1\}$, such that $\mathsf{smCE}(\Gamma_\varepsilon) \ge \Omega(\varepsilon^{O(1)})$, and $\mathsf{kCE}^{\mathsf{Gauss}}(\Gamma_\varepsilon) \le O(\exp(-1/\varepsilon))$, where $\mathsf{kCE}^{\mathsf{Gauss}} := \mathsf{kCE}^{K_{\mathsf{Gauss}}}$ is the Gaussian kernel calibration error with $K_{\mathsf{Gauss}}(u, v) = \exp(-(u - v)^2)$.*

ACKNOWLEDGMENTS

Part of this work was performed while LH was interning at Apple. LH is also supported by Moses Charikar's and Omer Reingold's Simons Investigators awards, Omer Reingold's NSF Award IIS-1908774, and the Simons Foundation Collaboration on the Theory of Algorithmic Fairness.

JB is supported by a Junior Fellowship from the Simons Society of Fellows.

PN and JB acknowledge the city of New Orleans, for providing an environment conducive to both research and recreation in the early stages of this project. PN also acknowledges his partner Shreya Shankar for their invaluable support.

REFERENCES

[1] Imanol Arrieta-Ibarra, Paman Gujral, Jonathan Tannen, Mark Tygert, and Cherie Xu. 2022. Metrics of calibration for probabilistic predictions. *arXiv preprint arXiv:2205.09680* (2022).

[2] A. Berlinet and C. Thomas-Agnan. 2011. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Springer US. https://books.google.com/books?id=bX3TBwAAQBAJ

[3] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. 1990. Self-Testing/Correcting with Applications to Numerical Problems. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, May 13-17, 1990, Baltimore, Maryland, USA*, Harriet Ortiz (Ed.). ACM, 73–83.

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

[5] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. 2022. A Unifying Theory of Distance from Calibration. arXiv:2211.16886 [cs.LG]

[6] Cynthia S Crowson, Elizabeth J Atkinson, and Terry M Therneau. 2016. Assessing calibration of prognostic risk scores. *Statistical methods in medical research* 25, 4 (2016), 1692–1706.

[7] A. P. Dawid. 1982. Objective Probability Forecasts. *University College London, Dept. of Statistical Science. Research Report 14* (1982).

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

[9] A Philip Dawid. 1984. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)* 147, 2 (1984), 278–290.

[10] A Philip Dawid. 1985. Calibration-based empirical probability. *The Annals of Statistics* (1985), 1251–1274.

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

[12] Kunio Doi. 2007. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics* 31, 4-5 (2007), 198–211.

[13] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. 2021. Outcome Indistinguishability. In *ACM Symposium on Theory of Computing (STOC'21)*. https://arxiv.org/abs/2011.13426

[14] Dean P. Foster and Sergiu Hart. 2018. Smooth calibration, leaky forecasts, finite recall, and Nash dynamics. *Games Econ. Behav.* 109 (2018), 271–293. https://doi.org/10.1016/j.geb.2017.12.022

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

[16] Oded Goldreich, Shafi Goldwasser, and Dana Ron. 1998. Property Testing and its Connection to Learning and Approximation. *J. ACM* 45, 4 (1998), 653–750.

[17] Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. 2023. Loss Minimization Through the Lens Of Outcome Indistinguishability. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 251)*, Yael Tauman Kalai (Ed.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 60:1–60:20. https://doi.org/10.4230/LIPIcs.ITCS.2023.60

[18] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. 2022. Omnipredictors. In *Innovations in Theoretical Computer Science (ITCS'2022)*. https://arxiv.org/abs/2109.05389

[19] Parikshit Gopalan, Michael P. Kim, Mihir Singhal, and Shengjia Zhao. 2022. Low-Degree Multicalibration. In *Conference on Learning Theory, 2-5 July 2022, London, UK (Proceedings of Machine Learning Research, Vol. 178)*. PMLR, 3193–3234.

[20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 1321–1330.

[21] Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. 2020. Calibration of neural networks using splines. *arXiv preprint arXiv:2006.12800* (2020).

[22] Cleve Hallenbeck. 1920. Forecasting Precipitation in Percentages of Probability. *Monthly Weather Review* 48, 11 (1920), 645–647.

[23] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML*.

[24] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. 2012. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association* 19, 2 (2012), 263–274.

[25] Sham Kakade and Dean Foster. 2008. Deterministic calibration and Nash equilibrium. *J. Comput. System Sci.* 74(1) (2008), 115–130.

[26] Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. 2021. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 29768–29779.

[27] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144* (2017).

[28] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. 2564–2572.

[29] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*.

[30] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine* 4, 1 (2021), 1–6.

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

[32] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. 2018. Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 2805–2814. https://proceedings.mlr.press/v80/kumar18a.html

[33] Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. 2022. T-Cal: An optimal test for the calibration of predictive models. *arXiv preprint arXiv:2203.01850* (2022).

[34] Feng Li, Masahito Aoyama, Junji Shiraishi, Hiroyuki Abe, Qiang Li, Kenji Suzuki, Roger Engelmann, Shusuke Sone, Heber MacMahon, and Kunio Doi. 2004. Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy. *American Journal of Roentgenology* 183, 5 (2004), 1209–1215.

[35] Matthew Mealiffe, Renee Stokowski, Brian Rhees, Ross Prentice, Mary Pettinger, and David Hinds. 2010. Clinical validity assessment of a breast cancer risk model combining genetic and clinical information. *Nature Precedings* (2010), 1–1.

[36] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 15682–15694.

[37] Allan H Murphy. 1998. The early history of probability forecasts: Some extensions and clarifications. *Weather and forecasting* 13, 1 (1998), 5–15.

[38] Allan H Murphy and Robert L Winkler. 1984. Probability forecasting in meteorology. *J. Amer. Statist. Assoc.* 79, 387 (1984), 489–500.

[39] Basil Mustafa, Aaron Loh, Jan Freyberg, Patricia MacWilliams, Megan Wilson, Scott Mayer McKinney, Marcin Sieniek, Jim Winkens, Yuan Liu, Peggy Bui, et al. 2021. Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913* (2021).

[40] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. 2014. Binary classifier calibration: Non-parametric approach. *arXiv preprint arXiv:1401.3390* (2014).

[41] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, Vol. 2015. NIH Public Access, 2901.

[42] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring Calibration in Deep Learning.. In *CVPR Workshops*, Vol. 2.

[43] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. 2006. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.* 72, 6 (2006), 1012–1042.

[44] Rahul Rahaman et al. 2021. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems* 34 (2021), 20063–20075.

[45] Ali Rahimi and Benjamin Recht. 2007. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf

[46] Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. 2022. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4036–4054.

[47] Ronitt Rubinfeld and Madhu Sudan. 1996. Robust Characterizations of Polynomials with Applications to Program Testing. *SIAM J. Comput.* 25, 2 (1996), 252–271.

[48] Man-hung Siu, Herbert Gish, and Fred Richardson. 1997. Improved estimation, evaluation and applications of confidence measures for speech recognition. In *Fifth European Conference on Speech Communication and Technology*.

[49] Ben Van Calster and Andrew J Vickers. 2015. Calibration of risk prediction models: impact on decision-analytic performance. *Medical decision making* 35, 2 (2015), 162–169.

[50] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*. PMLR, 11117–11128.

Received 2022-11-07; accepted 2023-02-06