Calibration through the Lens of Indistinguishability

PARIKSHIT GOPALAN
Apple
and
LUNJIA HU
Harvard University

Calibration is a classical notion from the forecasting literature which aims to address the question: how should predicted probabilities be interpreted? In a world where we only get to observe (discrete) outcomes, how should we evaluate a predictor that hypothesizes (continuous) probabilities over possible outcomes? The study of calibration has seen a surge of recent interest, given the ubiquity of probabilistic predictions in machine learning. This survey describes recent work on the foundational questions of how to define and measure calibration error, and what these measures mean for downstream decision makers who wish to use the predictions to make decisions. A unifying viewpoint that emerges is that of calibration as a form of indistinguishability, between the world hypothesized by the predictor and the real world (governed by nature or the Bayes optimal predictor). In this view, various calibration measures quantify the extent to which the two worlds can be told apart by certain classes of distinguishers or statistical measures.

Categories and Subject Descriptors: G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms: Measurement, Reliability, Theory

Additional Key Words and Phrases: Calibration, Uncertainty quantification, Prediction, Decision making

1. INTRODUCTION

Prediction is arguably the ubiquitous computational task of our time. Every day, a remarkable amount of computational resources are invested in the prediction of various probabilities, whether it is a language model trying to answer a user's ambiguous query or a recommendation engine trying to predict which product/profile to show a user. These automated predictions affect nearly every aspect of our lives, be it social, medical or financial. What makes prediction different from more classical computational tasks (such as sorting numbers or computing max-flows) is that there is no well-defined notion of what constitutes correctness.

To explore this issue in greater detail, let us consider the simplified setting of binary prediction, where nature is modeled as a joint distribution $\mathcal{D}^*$ over attributes $\mathbf{x}$ drawn from a domain $\mathcal{X}$ and labels $\mathbf{y} \in \mathcal{Y}$. In this article, we will mainly focus on the setting $\mathcal{Y} = \{0, 1\}$ of Boolean labels.[1] We denote the marginal distribution over $\mathcal{X}$ by $\mathcal{D}^*_{\mathcal{X}}$, and $\mathcal{Y}$ by $\mathcal{D}^*_{\mathcal{Y}}$. A predictor is a function $p : \mathcal{X} \to [0, 1]$. The *ground truth* in this setting is represented by the Bayes optimal predictor $p^*(x) = \mathbb{E}[\mathbf{y}^*|\mathbf{x} = x]$.

[1]We use boldface for random variables, thus $\mathbf{x}$ is random variable drawn from $\mathcal{X}$ whereas $x \in \mathcal{X}$ is a point in the domain.

Authors' addresses: parikg@apple.com, lunjia@alumni.stanford.edu

The obvious formulation of correctness in prediction might be to learn p^*. The challenge is that we never see the values of p^* itself, our only access to it is via the labels $\mathbf{y}^*$ which satisfy $\mathbb{E}[\mathbf{y}^*|x] = p^*(x)$. So the obvious formulation of correctness, as finding p which is close to p^* under some suitable measure of distance, will not work. There are (at least) two different and complementary approaches: loss minimization and calibration.

1.1 Loss minimization

In loss minimization, we choose a loss function $\ell : \{0,1\} \times [0,1] \to \mathbb{R}$, and a hypothesis class of predictors $\mathcal{P} = \{p : \mathcal{X} \to [0,1]\}$, and aim to find the predictor that minimizes

$$p = \arg\min_{p' \in \mathcal{C}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}^*} [\ell(\mathbf{y}^*, p'(\mathbf{x}))].$$

In essence, we use the labels $\mathbf{y}^*$ as a proxy for p^*, while ℓ plays the role of a distance measure. But it turns out that (for any proper loss) we indeed find the predictor in our family $\mathcal{P}$ that is closest to p^*. This is a consequence of the *bias variance* decomposition. Taking the example of squared loss $\ell(y,v) = (y-v)^2$, the decomposition tells us that for any predictor p, [2]

$$\mathop{\mathbb{E}}_{\mathcal{D}^*} [(\mathbf{y}^* - p(\mathbf{x}))^2] = \underbrace{\mathop{\mathbb{E}}_{\mathcal{D}^*} [(p(\mathbf{x}) - p^*(\mathbf{x}))^2]}_{\text{bias}} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{D}^*} [(\mathbf{y}^* - p^*(\mathbf{x}))^2]}_{\text{variance}}.$$

Note that the variance is a property of p^*, independent of p.

Loss minimization is a simple yet immensely powerful paradigm that powers much of contemporary machine learning. But is it a satisfactory notion of correctness for prediction tasks? Here are some questions to consider:

—Imagine that a decision maker is using a predictor to make decisions that minimize their own loss function. This loss may differ from the one used to train the model, and might differ across various decision makers. For instance, we could use forecasts about rain to decide whether or not to carry an umbrella, to decide whether to have a party outdoors or indoors, or whether to turn off the sprinklers. Each of these has its own loss function. Say our loss for carrying an umbrella when it does not rain is 0.1, and for not carrying an umbrella when it rains is 0.9. The optimal strategy here is to carry an umbrella on days when $p^*(x) \geq 0.1$. Now suppose that the predictor p we have access to is not Bayes optimal. How do we make decisions using this predictor? Should we carry an umbrella whenever $p(x) \geq 0.1$, just like with the Bayes optimal predictor, or should we make decisions differently?

—We know that the squared loss decomposes into bias and variance, but we have no way of knowing how large each of these are. If we suffer large squared loss, it could because nature is inherently random (e.g p^* is often close to $1/2$), or because nature is deterministic but sufficiently complex that it *looks random* to

[2] It is easy to prove a similar statement about any *proper* loss, and a little harder to prove it about arbitrary losses. But the takeaway remains the same: by minimizing loss over a family $\mathcal{P}$, we find the best approximation to p^* from $\mathcal{P}$ under a suitable notion of distance tailored to the loss.

our hypothesis class $\mathcal{P}$. Loss minimization does not distinguish between these scenarios.

—Suppose we wish to predict the probability of rain tomorrow, and the model p found by minimizing squared loss gives a 60% chance of rain. How should we interpret this prediction? Is it possible that although p minimizes expected error globally, it is not particularly good at prediction for certain types of days (like days in September)? Concerns like these arise naturally in the context of fair predictions for subgroups (see the discussion on multicalibration in Section 1.4).

The question of what a prediction really guarantees naturally leads us to calibration.

1.2 Calibration

Calibration is a notion of correctness that focuses on ensuring that predicted probabilities align with actual outcomes. Intuitively, on days when a calibrated predictor predicts a 60% chance of rain, it rains 60% of the time. Formally, we can define perfect calibration as follows:

Definition 1.1. The predictor $p : \mathcal{X} \to [0,1]$ is *perfectly calibrated* under the distribution $\mathcal{D}^*$ if for every $v \in \mathsf{Image}(p)$, it holds that $\mathbb{E}[\mathbf{y}^*|p(\mathbf{x}) = v] = v$.

A key property of calibration is that it simplifies downstream decision making. For instance, let us return to the problem of using the forecast about rain to decide whether or not to carry an umbrella, where our loss for carrying an umbrella when it does not rain is 0.1, and for not carrying an umbrella when it rains is 0.9. Now suppose that the predictor p we have access to is not Bayes optimal, but it is calibrated. If we are basing decisions solely on p, then the optimal strategy is still to carry an umbrella on days when $p(x) \geq 0.1$. The expected loss we would suffer is exactly what we would have suffered if our predictor were Bayes optimal.

This naturally motivates an alternate view of calibration as a notion of correctness for predictors based on indistinguishability from the Bayes optimal, which will be an important theme in this survey. This view is inspired by the outcome indistinguishability framework of [Dwork et al. 2021].[3]

To every predictor $p : \mathcal{X} \to [0,1]$, we can associate a distribution $\mathcal{D}^p$ on pairs $(\mathbf{x}, \mathbf{y}^p)$ where the marginal on $\mathbf{x}$ is $\mathcal{D}^*_{\mathcal{X}}$ and where $\mathbb{E}[\mathbf{y}^p|\mathbf{x}] = p(\mathbf{x})$. The Bayes optimal predictor for $\mathcal{D}^p$ is p. Perfect Calibration requires that the joint distributions $(p(\mathbf{x}), \mathbf{y}^p)$ and $(p(\mathbf{x}), \mathbf{y}^*)$ be identical.

Lemma 1.2 Perfect calibration as indistinguishability. *The predictor $p : \mathcal{X} \to [0,1]$ is perfectly calibrated under the distribution $\mathcal{D}^*$ iff the joint distributions $J^* = ((p(\mathbf{x}), \mathbf{y}^*))$ and $J^p = (p(\mathbf{x}), \mathbf{y}^p)$ on $[0,1] \times \{0,1\}$ are identical.*

Let us see why this is true. Since the marginal distribution of $\mathbf{x}$ is the same in both cases, the distribution of $p(\mathbf{x})$ is also the same. In essence, we require that the distributions $\mathbf{y}^*|p(\mathbf{x})$ and $\mathbf{y}^p|p(\mathbf{x})$ be identical. Since the latter is the Bernoulli distribution with parameter $p(\mathbf{x})$, we require the same for $\mathbf{y}^*|p(\mathbf{x})$, which

[3]That work does not consider calibration *per se*, it instead considers more general notions such as multicalibration from [Hébert-Johnson et al. 2018]. In the context of calibration, it is plausible that this indistinguishability viewpoint predates it, though we have not found a reference.

is the standard definition. This guarantee conditional on each prediction is the key strength of calibration as a prediction guarantee.[4]

The indistinguishability property asserts that $p(\mathbf{x})$ be a plausible explanation for the observations $\mathbf{y}^*$ given $\mathbf{x}$, in that the conditional distribution of $\mathbf{y}^*|p(\mathbf{x})$ is consistent with the hypothesis that $p(\mathbf{x})$ is the Bayes optimal predictor. This indistinguishability property is desirable in machine learning, where we often try to model complicated processes (like the likelihood of a medical condition) and are unlikely to find the true Bayes optimal. Calibrated predictors are considered more trustworthy, whereas a predictor that is not calibrated will fail some basic tests: the probability of the label being 1 conditioned on $p(\mathbf{x}) = v$ is not v.

In the bigger picture, the notion of indistinguishability has played a central role in several disciplines within theoretical computer science, cryptography and pseudorandomness to name just a couple, indeed its roots go back to the Turing test. Viewing calibration as a form of indistinguishability lets us draw on ideas from those areas when we seek to define approximate calibration or generalize our notions beyond the binary classification setting.

1.3 From perfect to approximate calibration

Perfect calibration is a clean abstraction, but predictors trained and used for prediction tasks in the real world are seldom perfectly calibrated. For calibration to be a useful notion, we need to define what it means for a predictor to be approximately (but not perfectly) calibrated, and we need efficient methods to measure calibration error. How to do this in a principled manner is the main focus of this article.

There are many desiderata that one might hope a notion of approximate calibration satisfies:

(1) It should preserve the desirable properties of calibration, such as indistinguishability and simple downstream decision making, in some approximate sense.

(2) It should be efficient to measure the calibration error of a given predictor, just from black box access to samples of the form $(p(\mathbf{x}), \mathbf{y}^*)$, both in terms of sample complexity and computational complexity. In an online setting (to be defined shortly), we might wish for our notion to have low-regret algorithms.

(3) The notion should be robust to small perturbations in the predictor. A tiny change to a calibrated predictor should not result in a predictor with huge calibration error. For instance, changing the days forecast from 60% to 59.999% should not result in wild swings in the calibration error.[5]

(4) The notion should extend beyond binary classification, to multiclass labeling and regression, while maintaining properties like efficiency.

Achieving all of these properties is not easy. The classical notion of calibration error, which is the expected calibration error or ECE, only satisfies property (1) above; we will discuss this in more detail in Section 3. An active line of recent

[4]Of course, there might be a different calibrated predictor that only puts the chances of rain at 30%. There is no contradiction because the level sets of the predictors over which we average are different in the two cases.

[5]This is especially desirable from a machine learning perspective, where the lower order bits of prediction are considered insignificant and typically disregarded in low-precision arithmetic.

research has yielded a rich theory of approximate notions of calibration, together with algorithms for computing them efficiently in various models. Yet, to date, there is no single notion that satisfies all four desiderata mentioned above!

Perhaps this is too much to hope for, since some of these desiderata (eg. robustness and low-regret algorithms) arise from different motivating scenarios. But a clear takeaway from this body of research is that approximate calibration is surprisingly challenging to define and measure. The key technicality in defining approximate calibration error comes from conditioning. Every definition of calibration involves some form of conditioning on predictions. While this conditioning is simple for perfectly calibrated predictors, it is far trickier for predictors that are not perfectly calibrated, since predictions are real-valued.

In this survey, we will highlight how the *indistinguishabilty* viewpoint on calibration guides us in formulating what approximate calibration should mean. At a high level, there are two approaches to this task:

—**Limit the set of distinguishers :** Rather than require J^* and J^p be identical, we ask that they look similar to a family W of distinguishers. The calibration error is measured by the maximum distinguishing advantage achieved over all distinguishers in W. This approach is directly inspired by cryptography and pseudorandomness.

—**Use a divergence/distance on distributions:** Since J^* and J^p are both distributions on the domain $[0,1]\times\{0,1\}$, we can use distance measures/divergences on probability distributions (e.g., total variation, earthmover) to measure the distance between them, and use this as our measure of the calibration error. As we will see, this view relates to a quantification of the economic value of calibration from the perspective of downstream decision making.

These approaches lead to a number of calibration error measures that we will explore in more detail in this article, and which have many advantages over ECE and other traditional calibration measures. We will analyze smooth calibration error [Kakade and Foster 2008], which satisfies properties (1-3) but not (4). It also corresponds to an intuitive notion of approximate calibration, where the predictor is close to some perfectly calibrated predictor in earthmover distance.

From the computational standpoint, the natural model in which to study calibration has been the online setting, where we measure the regret or calibration error of our prediction strategy over T time steps.[6] The classic work of [Foster and Vohra 1998] showed that sublinear calibration error, as measured by ECE is possible. The regret rate achieved in their work is $O(T^{2/3})$.[7] It is known that regret rates of $O(\sqrt{T})$ or even $\widetilde{O}(\sqrt{T})$ are not possible for ECE [Qiao and Valiant 2021], and figuring out the optimal regret achievable is an active area of research (see, e.g., [Dagan et al. 2025]). However, new notions of calibration, which we will discuss in this survey, actually admit prediction strategies that achieve $O(\sqrt{T})$ or $\widetilde{O}(\sqrt{T})$ regret rates [Qiao and Zheng 2024; Arunachaleswaran et al. 2025; Hu and Wu 2024].

[6]Note that the computational task of learning a calibrated predictor admits trivial solutions in the offline model; for instance, one can always predict the expectation of the label.

[7]The regret rate is T times the calibration error on the uniform distribution over the T time steps.

1.4 Limitations and generalizations

Calibration is clearly a desirable property for a predictor, but it has limitations, and cannot be considered as a standalone notion of goodness for a predictor. We ideally want predictors to have both good calibration and other properties like small expected loss. We discuss these limitations below, and use this as motivation to introduce the stronger notion of multicalibration [Hébert-Johnson et al. 2018], and discuss how it addresses these limitations.

Calibration does not guarantee utility. There are many predictors that will satisfy calibration, and we would not consider all of them to be equally informative or good. For instance, the *average* predictor $\bar{p}$ that always predicts the average label $\mathbb{E}_{\mathcal{D}^*}[\mathbf{y}^*]$ is perfectly calibrated, as is the Bayes optimal p^*. Any reasonable loss function would distinguish between these predictors, but calibration (by itself) does not.

Calibration gives guarantees on average over the entire population. In some applications, this might not be good enough. For instance, suppose we train a predictor to predict the risk of a certain risk of disease for a patient. On examining the data, we find that although the predictor is calibrated over the general population, it is miscalibrated for patients with a certain medical history, who are a small fraction of the dataset (so this does not affect the overall calibration error too much). We would not trust such a predictor to make decisions for those patients.

Multicalibration. Multicalibration, introduced in [Hébert-Johnson et al. 2018], is a strengthening of calibration. It requires that our predictions are calibrated, even when conditioned on membership in a rich collection of demographic subgroups $\mathcal{C} \subseteq 2^{\mathcal{X}}$. Which subgroups to consider is an important consideration, which is dictated by the data and computational resources available to the predictor. We refer the reader to [Hébert-Johnson et al. 2018] for more details.

Although calibration by itself does not guarantee good loss minimization, multicalibration with respect to rich class of subgroups $\mathcal{C}$ does imply strong loss minimization. This was the key insight in the work of [Gopalan et al. 2022] which introduced the notion of omniprediction. Omniprediction asks for a predictor that is as good as benchmark class $\mathcal{C}$ not just for a single loss function, but for any loss from a large family of loss functions. [Gopalan et al. 2022] shows a surprising connection between omniprediction with respect to a benchmark class $\mathcal{C}$ and multicalibration with respect to $\mathcal{C}$.

From the indistinguishability perspective, [Dwork et al. 2021] showed that multicalibration is equivalent to indistinguishability of the distributions $(c(\mathbf{x}), p(\mathbf{x}), \mathbf{y}^*)$ and $(c(\mathbf{x}), p(\mathbf{x}), \mathbf{y}^p)$ for all $c : \mathcal{X} \to \{0, 1\}$ that lie in some family $\mathcal{C}$ of functions. Beyond its original motivation in multigroup fairness, multicalibration has proved to be tremendously powerful, finding applications to omniprediction [Gopalan et al. 2022], domain adaptation [Kim et al. 2022], pseudorandomness [Dwork et al. 2023], and computational complexity [Casacuberta et al. 2024].

Organization of this survey. In Section 2, we consider expected calibration error (ECE) and explore its weaknesses. In Section 3, we introduce weighted calibration measures which capture the notion of indistinguishability to limited classes of distinguishers. This unifies several different notions of approximate calibration

in the literature. In Section 4, we describe Calibration decision loss, which looks at calibration from an economics perspective, through the eyes of a downstream decision maker who wants to use the predictions of a predictor to optimize their utility. We review the active area of research on online calibration in Section 5. Given the number of calibration notions that we will encounter, a natural question is whether there is some ground truth notion against which we can compare these different notions. In Section 6, we define the distance to calibration, which proposes a ground-truth notion of what approximate calibration ought to mean, and show how smooth calibration shows up naturally in this setting. In the interest of brevity, we omit most proofs from the survey. We direct the interested reader to the arXiv for a fuller version of this article that includes full proofs and some additional material.

2. EXPECTED CALIBRATION ERROR

We start with what is arguably the most popular metric for measuring calibration error: the expected calibration error or ECE. We examine some of its shortcomings, which will guide us in formulating other notions of approximate calibration.

Definition 2.1. The *expected calibration error* of a predictor p under $\mathcal{D}^*$ is defined as $\mathrm{ECE}(p, \mathcal{D}^*) = \mathbb{E}\,|\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})] - p(\mathbf{x})]|$.

Some notes on the definition of ECE:

—While perfect calibration requires $\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})] = p(\mathbf{x})$, ECE allows for some slack in the equality, and measures the average deviation over all p.

—We have defined ECE as measuring the absolute deviation between $\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})]$ and $p(\mathbf{x})$. We could instead have used the square or the q^{th} power for $q \geq 1$ and defined $\mathrm{ECE}_q(p, \mathcal{D}^*) = \mathbb{E}[|\,\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})] - p(\mathbf{x})|^q]^{1/q}$. By the convexity of t^q, ECE_q is an increasing function of q.

For a better understanding of ECE, we look at two alternative characterizations. The first characterizes it in terms of the maximum inner product with a distinguisher b which is a bounded function on $[0, 1]$.

LEMMA 2.2. *Let* $B = \{b : \{0, 1\} \to [-1, 1]\}$ *be the family of all bounded functions. Then* $\mathrm{ECE}(p, \mathcal{D}^*) = \max_{b \in B} \mathbb{E}_{J^*}[b(x)(\mathbf{y}^* - p(\mathbf{x}))]$.

For two distributions $\mathcal{D}_1, \mathcal{D}_2$ on a domain $\mathcal{X}$, we define

$$\mathrm{TV}(\mathcal{D}_1, \mathcal{D}_2) = \max_{S \subseteq \mathcal{X}} |\mathcal{D}_1(S) - \mathcal{D}_2(S)|.^8$$

We state the second characterization in terms of total variation distance.

LEMMA 2.3. *We have* $\mathrm{ECE}(p, \mathcal{D}^*) = \mathrm{TV}(J^*, J^p)$.

The trouble with ECE. At first glance, ECE seems to be a reasonable measure of calibration error. However there are (at least) a couple of problems with it: it is hard to efficiently estimate (even in the binary classification setting), and it is very discontinuous. Thus it fails desiderata (2-4).

[8]When the space $\mathcal{X}$ is infinite, we must restrict S to be measureable, but we will ignore this and other such subtleties.

The computational difficulty stems from Lemma 2.2. Estimating the ECE is equivalent to finding the best witness $b \in B$. This is essentially a learning problem over a class with infinite VC dimension. Indeed, one can show that sample complexity of estimating ECE can be as large as $\Omega(\sqrt{|\mathcal{X}|})$. Ideally, we would like to complexity to be independent of the domain size, and depending only on the desired estimation error.

The continuity problems are hinted at by Lemma 2.3. While total variation distance is a good distance measure for distributions over discrete domains, it is not ideal for continuous domains. And our setting involving distributions over predictions in $[0, 1]$ is inherently continuous. As the next example illustrates, ECE turns out to be highly discontinuous in the predictions of our predictor.

—Let $\mathcal{D}_2$ be the uniform distribution a two point space $\{(a, 0), (b, 1)\}$, where a is always labeled 0 and b is labeled 1.

—Consider the predictor p_0 which predicts $1/2$ for both a and b. It is perfectly calibrated, hence $\mathrm{ECE}(p_0) = 0$.

—For $\epsilon > 0$, define the predictor p_ϵ where $p_\epsilon(a) = 1/2 - \epsilon, p_\epsilon(b) = 1/2 + \epsilon$. It is easy to verify that $\mathrm{ECE}(p_\epsilon) = 1/2 - \epsilon$.

Think of ϵ being infinitesimally small but positive, so that p_ϵ is extremely close to p_0. Intuitively, p_ϵ is very close to being perfectly calibrated, it only requires a small perturbation of the lower order bits. Yet, the ECE is close to $1/2$ for p_ϵ, whereas it is 0 for p_0.

There are many ad-hoc fixes in practice that aim to get around these difficulties. For instance, bucketed ECE divides the interval $[0, 1]$ into b equal sized buckets, rounds the predictions in each bucket (say to the midpoint) and then measures the ECE of the discretized predictor. But [Blasiok et al. 2023a] observe that this results in a bucketed ECE which oscillates between 0 and $1/2 - \epsilon$ depending on whether the number of buckets is odd or even!

Are our issues with ECE small technicalities or symptoms of a bigger problem? We believe it is the latter. Assume you are training a predictive model, and you measure its ECE and find it to be large. Is this something you should worry about? Is your model truly miscalibrated (whatever that means)? Or is there an infinitesimal perturbation of its predictions that will make it perfectly calibrated? In general, there are sound reasons to prefer metrics that are reasonably smooth. It is also important for estimation to be efficient in terms of both samples and computation, which is not the case for ECE.

3. WEIGHTED CALIBRATION ERROR

In this section, we will explore notions of approximate calibration that only require that J^* and J^p look similar to a family W of distinguishers or weight functions. This results in a general template called weighted calibration, which is parametrized by the family W. Instantiating this notion with the family of bounded Lipschitz functions, we derive the notion of smooth calibration [Kakade and Foster 2008]. We briefly describe some other notions of calibration from the literature that can be viewed as instantiations of this template.

3.1 Weighted calibration

Weighted calibration error [Gopalan et al. 2022] captures the extent to which a collection of distinguishing functions are able to distinguish J^* from J^p. Since J^* and J^p are both distributions over $[0,1] \times \{0,1\}$, we consider distinguishing functions $f : [0,1] \times \{0,1\} \to [-1,1]$. Since the second argument to f is Boolean, we can write $f(v,y) = w(v)y + u(v)$. Hence,

$$\begin{aligned}\mathop{\mathbb{E}}_{J^*}[f(\mathbf{v},\mathbf{y}^*)] - \mathop{\mathbb{E}}_{J^p}[f(\mathbf{v},\mathbf{y}^p)] &= \mathop{\mathbb{E}}_{J^*}[w(\mathbf{v})\mathbf{y}^*] - \mathop{\mathbb{E}}_{J^p}[w(\mathbf{v})\mathbf{y}^p] = \mathop{\mathbb{E}}_{J^*}[w(\mathbf{v})\mathbf{y}^*] - \mathop{\mathbb{E}}_{J^p}[w(\mathbf{v})\mathbf{v}] \\ &= \mathop{\mathbb{E}}_{J^*}[w(\mathbf{v})(\mathbf{y}^* - \mathbf{v})]. \end{aligned} \tag{3.1}$$

where the first and third equalities hold because $\mathbf{v}$ is identically distributed under J^* and J^p, and the second is because $\mathbb{E}[\mathbf{y}^p|\mathbf{v}] = \mathbf{v}$. This tells us that we can limit ourselves to distinguishers of the form $f(v,y) = w(v)y$, and the distinguishing advantage can be thought of as an expectation under the single distribution J^* (Equation (3.1)). This leads to the following definition from [Gopalan et al. 2022].

Definition 3.1 *Weighted calibration error [Gopalan et al. 2022].* Let $W = \{w : [0,1] \to [-1,1]\}$ be a family of weight functions. The *W-weighted calibration error* of the predictor $p : \mathcal{X} \to [0,1]$ is defined as

$$\mathrm{CE}_W(p, \mathcal{D}^*) = \max_{w \in W} \left| \mathop{\mathbb{E}}_{\mathcal{D}^*}[w(p(\mathbf{x}))(\mathbf{y}^* - p(\mathbf{x}))] \right|.$$

The definition of weighted calibration error suggests a natural computational problem: the problem of calibration auditing for a weight family W. This is the computational problem of deciding whether $\mathrm{CE}_W(p, \mathcal{D}^*)$ is 0 or exceeds α, given access to random samples $(p(\mathbf{x}), \mathbf{y}^*)$ from $\mathcal{D}^*$. This problem turns out to be closely related to agnostic learning for the class W, as shown by [Gopalan et al. 2024].

If we instantiate weighted calibration with $W = B$ where B is the set of all bounded functions introduced in 2.2, we recover ECE. But this also illustrates why ECE is hard to compute efficiently: the set B has infinite VC dimension, hence it cannot be learnt efficiently.

Note that we could have defined the weighted calibration error CE_W as a function of J^*, the joint distribution of $(p(\mathbf{x}), \mathbf{y}^*)$, rather than the pair $(p, \mathcal{D}^*)$. We prefer mentioning p explicitly for clarity, but it is important to note that CE_W only depends on J^*. Indeed, most common measures of calibration error and loss only depend on the distribution of J^*. For instance, the cross-entropy loss and square loss only depend on how labels and predictions are jointly distributed, not on whether we are labeling images or tabular data; if we predict $p(x) = 0.7$ and the label is 1, that fixes the loss suffered at x, regardless of the features x.

3.2 Smooth calibration

Smooth calibration, introduced by [Kakade and Foster 2008] is an instantiation of weighted calibration that restricts the class of weight functions to Lipschitz continuous functions. This ensures that small perturbations of the predictor do not result in large changes in the calibration error.

Definition 3.2. Let $L = \{l : [0,1] \to [-1,1]\}$ denote the subset of 1-Lipschitz functions from B. Define the *smooth calibration error* of the predictor p under the

distribution $\mathcal{D}^*$ as $\mathrm{smCE}(p, \mathcal{D}^*) = \mathrm{CE}_L(p, \mathcal{D}^*)$.

By only allowing Lipschitz weight functions, Smooth calibration ensures that the calibration error does not change dramatically under small perturbations of the predictor.[9] Given predictors $p_1, p_2 : \mathcal{X} \to [0,1]$ and a distribution $\mathcal{D}^*$ on $\mathcal{X}$, let the expected ℓ_1 distance between them be

$$d(p_1, p_2) = \mathop{\mathbb{E}}_{\mathcal{D}^*}[|p_1(\mathbf{x}) - p_2(\mathbf{x})|].$$

Smooth calibration error is Lipschitz in this distance.

LEMMA 3.3. *For any weight family $W \subseteq L$, $\mathrm{CE}_W(p, \mathcal{D}^*)$ is 4-Lipschitz in d.*

Returning to the example above with p_0 and p_ϵ, restricting to Lipschitz distinguishers means that smooth calibration considers p_ϵ to also be well calibrated, since its smooth calibration error is $O(\epsilon)$.

An alternate view of smooth calibration is in terms of earthmover distance between J^* and J^p. Consider the ℓ_1 metric on $[0,1] \times \{0,1\}$ where $\ell_1((v,y),(v',y')) = |v - v'| + |y - y'|$. For two distributions J, J' on $[0,1] \times \{0,1\}$, we denote the earthmover distance between two distributions under the ℓ_1 metric as $\mathrm{EMD}(J, J')$. Smooth calibration captures the earth-mover distance between J^* and J^p.

LEMMA 3.4. *We have* $\mathrm{EMD}(J^*, J^p)/2 \leq \mathrm{smCE}(p, \mathcal{D}^*) \leq \mathrm{EMD}(J^*, J^p)$.

This lemma should be contrasted with Lemma 2.3, which characterizes ECE in terms of the total variation distance.

We have defined smooth calibration error in terms of the family of 1-Lipschitz distinguishers. But since an L-Lipschitz function for $L > 1$ can be made 1-Lipschitz by rescaling the range by L, the calibration error can only increase by L even if we allow L-Lipschitz distinguishers.

3.3 Other notions of weighted calibration

We have seen two notions of weighted calibration so far: ECE and smCE. Several other calibration metrics that have been considered in the literature can be naturally viewed as instances of weighed calibration. We list some of them below.

—Low-degree calibration [Gopalan et al. 2022] corresponds to the case where $W = P_d$ consists of degree d polynomials. This class is fairly Lipschitz (since polynomials have bounded derivatives. The main attraction of this notion is that it is efficient to computer, even in the multiclass setting.

—In Kernel calibration [Kumar et al. 2018; Blasiok et al. 2023a] the family of weight functions lies in a Reproducing Kernel Hilbert Space. There are many choices of kernel possible, such as the Laplace kernel, the Gaussian kernel or the polynomial kernel, each of these results in distinct calibration measures with their own properties.

[9]Note that Lemma 2.2 tells us that there exists a bounded function b_ϵ that explains the high ECE for p_ϵ, specifically, $b_\epsilon(v) = \mathrm{sign}(v - 1/2)$. This function is discontinuous near 1/2, which causes the extreme sensitivity to perturbations.

4. CALIBRATION ERROR FOR DECISION MAKING

In this section, we will explore a second approach to relaxing the definition of perfect calibration, where rather than asking J^* and J^p be identical, we require them to be close when measured under a suitable divergence. This leads to another important measure of the calibration error, the Calibration Decision Loss (CDL), introduced recently by Hu and Wu [Hu and Wu 2024]. Underlying the notion of CDL is a concrete and natural quantification of the economic value of calibration from the perspective of downstream decision making.

We define the notion of CDL in Section 4.1 and discuss its alternative formulation using Bregman divergences between J^* and J^p in Section 4.3. A key tool we use to prove this Bregman divergence formulation is a classic characterization of *proper scoring rules* [McCarthy 1956; Savage 1971; Gneiting and Raftery 2007].

4.1 Calibration Decision Loss

The definition of the Calibration Decision Loss comes naturally when we look at calibration through an economic lens, from the perspective of downstream decision makers. What does calibration mean to a person who uses the predictions (e.g. chance of rain) to make downstream decisions (e.g. take an umbrella or not)? We will show that a calibrated predictor provides a concrete *trustworthiness* guarantee to *every* payoff-maximizing downstream decision maker (Theorem 4.1). This observation gives not only a characterization of perfect calibration, but also a natural way of quantifying the calibration error of a miscalibrated predictor, using the payoff loss caused by trusting the (miscalibrated) predictor in downstream decision making. This way of quantifying the calibration error leads exactly to Calibration Decision Loss (Definition 4.2).

We start by formally defining decision tasks. A decision task $\mathcal{T}$ has two components: an action space A and a payoff function $u : A \times \{0, 1\} \to \mathbb{R}$. Given a decision task $\mathcal{T} = (A, u)$, the decision maker must pick an action $a \in A$ in order to maximize the payoff $u(a, y) \in \mathbb{R}$. Here, the payoff depends not only on the chosen action a, but also on the true outcome $y \in \{0, 1\}$ unknown to the decision maker. For example, if the outcome $y \in \{0, 1\}$ represents whether or not it will be rainy today, a natural decision task may have two actions to choose from: $A = \{\text{take umbrella}, \text{not take umbrella}\}$. Each combination (a, y) of action and outcome corresponds to a payoff value $u(a, y)$ that may depend on the susceptibility to rain and the inconvenience of carrying an umbrella.

Prediction enables decision making under uncertainty. While the decision maker is unable to observe the true outcome y before choosing the action, we assume that they are assisted by a prediction $v \in [0, 1]$. In the ideal case, the prediction correctly represents the probability distribution of the true outcome. That is, the outcome y follows the Bernoulli distribution with parameter v (denoted $\mathbf{y} \sim v$). To maximize the expected payoff, the decision maker should choose the action

$$\sigma_{\mathcal{T}}(v) \in \arg\max_{a \in A} \mathop{\mathbb{E}}_{\mathbf{y} \sim v} u(a, \mathbf{y}) \tag{4.1}$$

in response to the (correct) prediction v. We call the function $\sigma_{\mathcal{T}} : [0, 1] \to A$ the *best-response function*. Throughout the section, we assume that each decision task $\mathcal{T} = (A, u)$ is associated with a well-defined best-response function. That is, we

focus on tasks $\mathcal{T}$ where the arg max in (4.1) is always non-empty.

In reality, predictions are seldom perfectly correct. It is thus unclear whether applying the best-response function would still lead to optimal payoff. The following theorem tells us that as long as the predictions are calibrated, the best response function remains the optimal mapping from predictions to actions, allowing the decision maker to *trust the predictions as if they were correct.*

THEOREM 4.1 CALIBRATED PREDICTIONS ARE TRUSTWORTHY. *Let $\mathcal{D}$ be a joint distribution on $\mathcal{X} \times \{0,1\}$. For any perfectly calibrated predictor $p : \mathcal{X} \to [0,1]$ and any decision task $\mathcal{T} = (A, u)$, it holds that*

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x})), \mathbf{y})] = \max_{\sigma:[0,1]\to A} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma(p(\mathbf{x})), \mathbf{y})]. \tag{4.2}$$

In other words, the maximum value of the expected payoff is attained when we choose $\sigma = \sigma_{\mathcal{T}}$. Conversely, if (4.2) holds for every decision task $\mathcal{T}$, then the predictor p is perfectly calibrated.

We defer the proof of Theorem 4.1 to Section 4.2 and discuss how it suggests a new calibration measure. According to the theorem, if a predictor p is miscalibrated, then the right-hand side of (4.2) is larger than the left-hand side for some decision task $\mathcal{T}$. The difference between the two sides is exactly the payoff loss incurred by the decision maker who follows the best-response strategy $\sigma_{\mathcal{T}}$ assuming (incorrectly) that the predictions were calibrated. Thus, a natural measure of the level of miscalibration is exactly this payoff loss. For a fixed decision task $\mathcal{T}$, this payoff loss is termed the *Calibration Fixed Decision Loss (CFDL)* [Hu and Wu 2024]. Taking the worst-case payoff loss over all decision tasks $\mathcal{T} = (A, u)$ with bounded payoff functions $u : A \to [0,1]$, we get the Calibration Decision Loss (CDL).

Definition 4.2 *Calibration Decision Loss (CDL) [Hu and Wu 2024].* Let $\mathcal{D}$ be a joint distribution over $\mathcal{X} \times \{0,1\}$. Given a predictor $p : \mathcal{X} \to [0,1]$, we define its *Calibration Fixed Decision Loss (CFDL)* with respect to a (fixed) decision task $\mathcal{T} = (A, u)$ as

$$\mathrm{CFDL}_{\mathcal{T}}(p, \mathcal{D}) := \max_{\sigma:[0,1]\to A} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma(p(\mathbf{x})), \mathbf{y})] - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x})), \mathbf{y})].$$

We define the *Calibration Decision Loss (CDL)* of the predictor p as the supremum of the CFDL over all decision tasks (A, u) where the payoff function $u : A \to [0,1]$ has its range bounded in $[0,1]$:

$$\mathrm{CDL}(p, \mathcal{D}) := \sup_{\mathcal{T}=(A,u), u:A\to[0,1]} \mathrm{CFDL}_{\mathcal{T}}(p, \mathcal{D}).$$

As we will see when we prove Theorem 4.1 in Section 4.2, the CDL is zero if and only if the predictor p is perfectly calibrated. If a predictor is not perfectly calibrated but has a small CDL, any decision maker can still trust the predictor as if it were calibrated without losing too much expected payoff. This holds because the CDL is the supremum of the CFDL over *all* payoff-bounded decision tasks.

We note that in the definition of CDL, decision tasks are restricted to have a bounded payoff function $u : A \to [0,1]$. This restriction is only for the purpose of normalization: multiplying the payoff function by any positive constant changes the corresponding CFDL by the same constant factor, whereas adding a constant

to the payoff function does not change the CFDL. There is no further restriction on the decision tasks beyond bounded payoff functions. In particular, the action set A can have arbitrary (even infinite) size. A small CDL implies that trusting the predictions will incur small payoff loss for *all* such decision tasks.

A natural question is how the CDL is related to other measures of the calibration error. We will prove that the CDL is quadratically related to the ECE:

THEOREM 4.3 [KLEINBERG ET AL. 2023; HU AND WU 2024]. *Let $\mathcal{D}$ be a joint distribution over $\mathcal{X} \times \{0,1\}$. For any predictor $p : \mathcal{X} \to [0,1]$,*

$$\mathrm{ECE}(p,\mathcal{D})^2 \le \mathrm{ECE}_2(p,\mathcal{D})^2 \le \mathrm{CDL}(p,\mathcal{D}) \le 2\,\mathrm{ECE}(p,\mathcal{D}) \le 2\,\mathrm{ECE}_2(p,\mathcal{D}). \quad (4.3)$$

Moreover, the quadratic relationship between CDL and ECE shown in Theorem 4.3 is tight (up to lower order terms): for any $\varepsilon \in (0, 1/10)$, there exist two pairs $(p_1,\mathcal{D}_1),(p_2,\mathcal{D}_2)$ such that

$$\mathrm{ECE}_2(p_1,\mathcal{D}_1) = \varepsilon, \qquad \mathrm{CDL}(p_1,\mathcal{D}_1) = 2\varepsilon;$$
$$\mathrm{ECE}(p_2,\mathcal{D}_2) = \varepsilon, \qquad \mathrm{CDL}(p_2,\mathcal{D}_2) \le \varepsilon^2 + O(\varepsilon^3).$$

We defer the proof of Theorem 4.3 to Section A.7. Here we briefly describe the two tight examples. The first example $(p_1,\mathcal{D}_1)$ is very simple. For $(\mathbf{x},\mathbf{y}) \sim \mathcal{D}_1$, we draw $\mathbf{y} \in \{0,1\}$ from the Bernoulli distribution with parameter $1/2+\varepsilon$, independent of $\mathbf{x}$. The predictor p_1 is the constant predictor $p_1(x) = 1/2$. In the second example, we draw $\mathbf{x}$ uniformly at random from the interval $[\varepsilon, 1]$ and then draw $\mathbf{y} \in \{0,1\}$ from the Bernoulli distribution with parameter $\mathbf{x} - \varepsilon$. The predictor p_2 is the identity function $p_2(x) = x$ for $x \in [\varepsilon, 1]$. We will prove the correctness of the examples in Section A.8.

The second example, $(p_2,\mathcal{D}_2)$, demonstrating that the CDL can be significantly smaller than the ECE, is quite instructive. It opens up the possibility that the CDL can be minimized at a faster rate than what is possible for ECE in the online setting. Indeed, the main technical result of [Hu and Wu 2024] gives an efficient online CDL minimization algorithm achieving rate $O(\sqrt{T}\log T)$, overcoming the information-theoretic lower bound $\Omega(T^{0.54389})$ for ECE [Qiao and Valiant 2021; Dagan et al. 2025] (see Section 5 for more discussions).

To conclude this subsection, CDL measures the calibration error using the payoff loss of downstream decision makers caused by mis-calibration. In addition to introducing CDL as a meaningful decision-theoretic measure of calibration, the work of [Hu and Wu 2024] also shows that CDL allows a significantly better rate than what is possible for ECE in online calibration, which we discuss in Section 5.

In Section 4.2 we give a simpler yet equivalent definition of the CFDL in (4.5), which leads to an interpretation of CDL through the lens of indistinguishability.

4.2 Characterization of the Maximum Expected Payoff

In this section we prove Theorem 4.1. We start by giving a characterization of the maximum expected payoff on the right-hand side of (4.2) for a general predictor p that may or may not be calibrated, which simplifies the definition of CFDL and will be useful in the proof.

Recall the definition of the recalibration $\widehat{p}$ of p (Definition 6.4): $\widehat{p}$ is obtained by replacing each prediction value $v = p(x)$ with the actual conditional expectation

$\mathbb{E}[\mathbf{y}|p(\mathbf{x}) = v]$. Clearly, $\widehat{p}$ is perfectly calibrated. If p is perfectly calibrated, then $\widehat{p} = p$. We have the following characterization of the maximum expected payoff achievable by post-processing p (see Section A.3 for proof):

LEMMA 4.4. *Let $\mathcal{D}$ be a joint distribution on $\mathcal{X} \times \{0,1\}$. For any predictor $p : \mathcal{X} \to [0,1]$ and any decision task $\mathcal{T} = (A, u)$, it holds that*

$$\max_{\sigma:[0,1]\to A} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma(p(\mathbf{x})), \mathbf{y})] = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(\widehat{p}(\mathbf{x})), \mathbf{y})], \tag{4.4}$$

where $\widehat{p}$ is the recalibration of p.

We can now rewrite the definition of CFDL (Definition 4.2) based on Lemma 4.4:

$$\mathrm{CFDL}_{\mathcal{T}}(p, \mathcal{D}) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(\widehat{p}(\mathbf{x})), \mathbf{y})] - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x})), \mathbf{y})]. \tag{4.5}$$

This expression allows us to easily calculate the CFDL for specific decision tasks. For example, consider the task $\mathcal{T}_2 = (A, u)$ where the action space A is the unit interval $A = [0,1]$, and the payoff function is quadratic:

$$u(a, y) = 1 - (a - y)^2 \in [0,1], \quad \text{for } a \in [0,1] \text{ and } y \in \{0,1\}.$$

The corresponding best-response function is the identity: $\sigma_{\mathcal{T}}(v) = v$. Plugging it in (4.5), we obtain an equality between the CFDL and the square of ECE_2:

$$\begin{aligned}
\mathrm{CFDL}_{\mathcal{T}_2}(p, \mathcal{D}) &= \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[(p(\mathbf{x}) - \mathbf{y})^2 - (\widehat{p}(\mathbf{x}) - \mathbf{y})^2] \\
&= \mathbb{E}[p(\mathbf{x})^2 - \widehat{p}(\mathbf{x})^2 + 2\mathbf{y}(\widehat{p}(\mathbf{x}) - p(\mathbf{x}))] \\
&= \mathbb{E}[p(\mathbf{x})^2 - \widehat{p}(\mathbf{x})^2 + 2\widehat{p}(\mathbf{x})(\widehat{p}(\mathbf{x}) - p(\mathbf{x}))] \quad (\mathbb{E}[\mathbf{y}|\widehat{p}(\mathbf{x}), p(\mathbf{x})] = \widehat{p}(\mathbf{x})) \\
&= \mathbb{E}[(p(\mathbf{x}) - \widehat{p}(\mathbf{x}))^2] = \mathrm{ECE}_2(p, \mathcal{D})^2.
\end{aligned} \tag{4.6}$$

We are now ready to prove Theorem 4.1.

PROOF OF THEOREM 4.1. If p is perfectly calibrated, then $p = \widehat{p}$, and (4.2) follows immediately from Lemma 4.4. For the reverse direction, if (4.2) holds for any decision task, then in particular, it holds for the task $\mathcal{T}_2$ above, implying $\mathrm{CFDL}_{\mathcal{T}_2}(p, \mathcal{D}) = 0$. By (4.6), we have $\mathrm{ECE}_2(p, \mathcal{D}) = 0$, so p is perfectly calibrated, as desired. Since the quadratic payoff function of $\mathcal{T}_2$ has a bounded range $[0,1]$, this proof also implies that the CDL of a predictor is zero if and only if the predictor is perfectly calibrated. □

4.3 The Bregman Divergence View of CDL

We show that the CFDL of a predictor p w.r.t. any decision task $\mathcal{T}$ can be expressed as a Bregman divergence $D_\varphi(J^* \| J^p)$ between the two joint distributions J^* and J^p (Theorem 4.9). Our proof uses a classic characterization of *proper scoring rules* [McCarthy 1956; Savage 1971; Gneiting and Raftery 2007].

We start with the definition of Bregman divergence.

Definition 4.5 *Bregman Divergence.* Let $\varphi : [0,1] \to \mathbb{R}$ be a convex function and let $\nabla\varphi : [0,1] \to \mathbb{R}$ be its subgradient. For any pair of values $\mu^*, \mu \in [0,1]$, their *Bregman divergence* w.r.t. φ is

$$D_\varphi(\mu^* \| \mu) := \varphi(\mu^*) - \varphi(\mu) - \nabla\varphi(\mu) \cdot (\mu^* - \mu).$$

Since $\nabla\varphi(\mu)$ is a subgradient of φ at μ, the Bregman divergence is always nonnegative. When $\mu = \mu^*$, the Bregman divergence becomes zero.

We will interpret the values $\mu^*, \mu \in [0,1]$ in the definition above as the parameters of two Bernoulli distributions. For example, if we choose $\varphi(\mu)$ to be the negative Shannon entropy of the Bernoulli distribution with parameter μ:

$$\varphi(\mu) = \mu \ln \mu - (1-\mu)\ln(1-\mu),$$

then the Bregman divergence becomes the KL divergence between the two Bernoulli distributions parameterized by μ^* and μ:

$$D_\varphi(\mu^* \| \mu) = \mu^* \ln \frac{\mu^*}{\mu} + (1-\mu^*) \ln \frac{1-\mu^*}{1-\mu}.$$

The following key theorem makes the connection between Bregman divergences and decision tasks.

Theorem 4.6. *For any decision task $\mathcal{T} = (A, u)$, there exists a convex function $\varphi : [0,1] \to \mathbb{R}$ with subgradient $\nabla\varphi : [0,1] \to \mathbb{R}$ such that*

$$u(\sigma_\mathcal{T}(v), y) = \varphi(v) + \nabla\varphi(v) \cdot (y - v) \quad \textit{for every } v \in [0,1] \textit{ and } y \in \{0,1\}.$$

To prove the theorem, one should first observe that the function $u(\sigma_\mathcal{T}(v), y)$ is a *proper scoring rule*. That is, for any $v, v' \in [0,1]$, we have

$$\mathop{\mathbb{E}}_{\mathbf{y}\sim v} u(\sigma_\mathcal{T}(v), \mathbf{y}) \geq \mathop{\mathbb{E}}_{\mathbf{y}\sim v} u(\sigma_\mathcal{T}(v'), \mathbf{y}),$$

which follows from the definition (4.1) of the best-response function $\sigma_\mathcal{T}$. The theorem then follows from a standard characterization of proper scoring rules [McCarthy 1956; Savage 1971; Gneiting and Raftery 2007].

We can now write the expected payoff achieved by a predictor p using the Bregman divergence between p and its recalibration $\widehat{p}$ (see Section A.4 for proof):

Lemma 4.7. *Fix a joint distribution $\mathcal{D}$ of $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \{0,1\}$. Let $p : \mathcal{X} \to [0,1]$ be a predictor and $\widehat{p}$ be its recalibration (Definition 6.4). Then for any decision task $\mathcal{T} = (A, u)$ and the corresponding convex function φ from Theorem 4.6,*

$$\mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_\mathcal{T}(p(\mathbf{x})), \mathbf{y})] = \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[D_\varphi(\widehat{p}(\mathbf{x}) \| p(\mathbf{x}))], \tag{4.7}$$

$$\mathrm{CFDL}_\mathcal{T}(p, \mathcal{D}) = \mathop{\mathbb{E}}_{\mathcal{D}}[D_\varphi(\widehat{p}(\mathbf{x}) \| p(\mathbf{x}))]. \tag{4.8}$$

We now generalize the definition of Bregman divergence to joint distributions, such as J^* and J^p, over the domain $[0,1] \times \{0,1\}$.

Definition 4.8 *Induced Bregman Divergence between Joint Distributions.* Let $\varphi : [0,1] \to \mathbb{R}$ be a convex function and let $\nabla\varphi : [0,1] \to \mathbb{R}$ be its subgradient. For any joint distribution J of $(\mathbf{v}, \mathbf{y}) \in [0,1] \times \{0,1\}$, we use $\mu_J(\mathbf{v}) = \mathbb{E}_J[\mathbf{y}|\mathbf{v}] \in [0,1]$ to denote the parameter of the Bernoulli distribution of $\mathbf{y}$ conditioned on $\mathbf{v}$. Let J_1, J_2 be a pair of joint distributions of $(\mathbf{v}, \mathbf{y}) \in [0,1] \times \{0,1\}$ that share the same marginal distribution of $\mathbf{v}$ and denote this marginal distribution by M. We define

the *Bregman divergence* between J_1 and J_2 *induced by* φ as[10]

$$D_\varphi(J_1 \| J_2) := \mathop{\mathbb{E}}_{\mathbf{v}\sim M}[D_\varphi(\mu_{J_1}(\mathbf{v}) \| \mu_{J_2}(\mathbf{v}))].$$

Combining Lemma 4.7 and Definition 4.8, we have a Bregman divergence characterization of the CFDL for any decision task $\mathcal{T}$ (see Section A.5 for proof).

THEOREM 4.9 BREGMAN DIVERGENCE VIEW OF CFDL. *Let $\mathcal{D}$ be a joint distribution over $\mathcal{X} \times \{0,1\}$, and let $p : \mathcal{X} \to [0,1]$ be a predictor. As before, given $(\mathbf{x}, \mathbf{y}^*) \sim \mathcal{D}$, we draw $\mathbf{y}^p$ from the Bernoulli distribution with parameter $p(\mathbf{x})$, and use J^*, J^p to denote the distributions of $(p(\mathbf{x}), \mathbf{y}^*)$ and $(p(\mathbf{x}), \mathbf{y}^p)$, respectively. Then for any decision task $\mathcal{T} = (A, u)$ and the corresponding convex function φ from Theorem 4.6,* $\mathrm{CFDL}_\mathcal{T}(p, \mathcal{D}) = D_\varphi(J^* \| J^p)$.

4.4 Further Work

As we discuss in this section, the defining property of CDL is that it provides a meaningful guarantee on the swap regret incurred by downstream decision makers who trust the predictions. However, CDL is undesirable in other aspects: like ECE, it is discontinuous and requires high sample complexity to estimate. Recent work of [Rossellini et al. 2025] introduces the notion of *cutoff calibration error* to address the sample complexity issue while maintaining a restricted form of the decision-theoretic guarantee of CDL (e.g. they consider the regret relative to *monotone* post-processings of the predictions). This notion of cutoff calibration is essentially identical to the notion of *proper calibration* from [Okoroafor et al. 2025], who give an algorithm achieving $\widetilde{O}(\sqrt{T})$ error rate for proper calibration in the online setting (see Section 5 for the setting). The works of [Blasiok et al. 2023b; Blasiok and Nakkiran 2024; Hartline et al. 2025] show that low smooth calibration error also gives certain decision-theoretic guarantees. In particular, these works show that it implies low regret for certain forms of Lipschitz post-processings or for decision makers who make randomized responses (e.g. by adding noise to the predictions), though this implication often comes with a quantitative loss (e.g. smooth calibration error being at most ε only implies an $O(\sqrt{\varepsilon})$ regret).

5. ONLINE CALIBRATION

We have discussed a variety of ways to quantify the calibration error of a given predictor. In this section, we turn to the algorithmic question of *constructing* a predictor with low calibration error. This question, when naively formulated, admits a trivial and unenlightening solution: one can simply construct a constant predictor that assigns (an approximation of) the overall average $\mathbb{E}[y]$ to every individual x. This is a well-calibrated predictor according to every calibration measure we have discussed. Thus, for the algorithmic question to be insightful, it is essential to formulate it in such a way that reaches beyond the trivial solution. The seminal work of Foster and Vohra [Foster and Vohra 1998] introduced one such interesting question that turned into an active area of research with exciting recent progress:

[10] One can also view $D_\varphi(J_1 \| J_2)$ as the Bregman divergence corresponding to the negative entropy $\Phi(J)$ of any joint distribution J of $(\mathbf{v}, \mathbf{y}) \in [0,1] \times \{0,1\}$ defined by $\Phi(J) := \mathbb{E}_{(\mathbf{v},\mathbf{y})\sim J}[\varphi(\mu_J(\mathbf{v}))]$.

calibration in *online* prediction. We will first describe the problem setting and then briefly survey some key results in the literature.

The online prediction problem has T rounds indexed by $t \in [T]$. In round t, our algorithm makes a prediction $p_t \in [0,1]$, and nature reveals an outcome $y_t \in \{0,1\}$. For example, we can interpret the problem as predicting the chance of rain each day for T days, where p_t is the prediction we make on day t, and $y_t = 1$ if day t is rainy. Since the rounds are ordered chronologically, we allow our algorithm to choose p_t as a function of the history $H_{t-1} = (p_1, \dots, p_{t-1}, y_1, \dots, y_{t-1})$, and similarly, y_t can depend on the history H_{t-1} as well.

To evaluate the calibration error of the prediction sequence $p_{1,\dots,T} := (p_1, \dots, p_T)$ w.r.t. the outcome sequence $y_{1,\dots,T} := (y_1, \dots, y_T)$, we consider the predictor $p : \{1, \dots, T\} \to [0,1]$ that assigns prediction $p(t) := p_t$ to each time step $t = 1, \dots, T$. Viewing each time step as an individual, we let $\mathcal{D}$ be the uniform distribution over the individual-outcome pairs (t, y_t) for $t = 1, \dots, T$. By slight abuse of notation, we can transform any calibration measure CAL for $(p, \mathcal{D})$ into a calibration measure CAL for $(p_{1,\dots,T}, y_{1,\dots,T})$ as follows:

$$\mathrm{CAL}(p_{1,\dots,T}, y_{1,\dots,T}) := T\,\mathrm{CAL}(p, \mathcal{D}).$$

Once a calibration measure CAL is chosen, our goal is to design a prediction algorithm that guarantees a small (e.g. sub-linear, i.e., $o(T)$) calibration error according to CAL, regardless of how the outcomes y_t are generated. We wish to design a prediction algorithm P that specifies how p_t should be chosen as a function of the history H_{t-1} for every round t. We want the calibration error to be small regardless of nature's strategy Y, which specifies how y_t should be chosen as a function of H_{t-1} for every round t. That is, we want to solve the following optimization problem:

$$\operatorname*{minimize}_{P} \max_{Y} \mathrm{CAL}(p_{1,\dots,T}, y_{1,\dots,T}), \text{ where } p_{1,\dots,T}, y_{1,\dots,T} \text{ is generated by } P \text{ and } Y.$$

For some calibration measures (e.g. ECE and CDL), it is necessary to use randomized prediction algorithms to achieve sub-linear rates. Such an algorithm constructs a distribution $\mathcal{P}$ over prediction strategies P to solve the following problem:

$$\operatorname*{minimize}_{\mathcal{P}} \max_{Y} \mathop{\mathbb{E}}_{P \sim \mathcal{P}}[\mathrm{CAL}(p_{1,\dots,T}, y_{1,\dots,T})].$$

Here is why randomized predictions are necessary for achieving sub-linear rates for ECE or CDL. For every deterministic prediction algorithm P, nature can infer the prediction p_t based on the history H_{t-1}, and can then choose $y_t = 1$ if and only if $p_t < 1/2$, incurring an $\Omega(T)$ rate for ECE and CDL.

In Table I, we summarize the current best upper and lower bounds on the optimal online calibration rates for a few calibration error measures we discussed earlier, which is an active topic for recent research. Notably, the only calibration measure in this table that does not allow an $\widetilde{O}(\sqrt{T})$ rate is ECE.

There are substantial gaps between the best upper and lower bounds for many calibration measures in this table, making it a natural question to close or reduce these gaps. Very recently, the works of [Peng 2025] and [Fishelson et al. 2025] have achieved significant progress on online calibration algorithms in the *multi-class* setting, opening up another exciting area for future research.

Calibration Error	Rate Upper Bound	Rate Lower Bound
Expected Calibration Error (ECE)	$O(T^{2/3})$ [Foster and Vohra 1998]	$\Omega(T^{1/2})$ [Folklore]
		$\Omega(T^{0.528})$ [Qiao and Valiant 2021]
	$O(T^{2/3-\varepsilon})$ [Dagan et al. 2025]	$\Omega(T^{0.54389})$ [Dagan et al. 2025]
Distance to Calibration [Blasiok et al. 2023a]	$O(T^{1/2})$ [Qiao and Zheng 2024] [Arunachaleswaran et al. 2025]	$\Omega(T^{1/3})$ [Qiao and Zheng 2024]
Smooth Calibration Error [Kakade and Foster 2008]	$O(T^{1/2})$ [Qiao and Zheng 2024] [Arunachaleswaran et al. 2025]	$\Omega(T^{1/3})$ [Qiao and Zheng 2024]
Calibration Decision Loss (CDL) [Hu and Wu 2024]	$O(T^{1/2}\log T)$ [Hu and Wu 2024]	$\Omega(T^{1/2})$ [Hu and Wu 2024]

Table I. Upper and lower bounds on the optimal rates for online calibration

6. THE DISTANCE TO CALIBRATION

At this point, we seem to have a Cambrian explosion of approximate calibration measures, each of which has their own desirable properties, and will give different calibration errors for a predictor. How should we compare these different measures, and decide which to use? Is there any notion of ground truth, that would guide us in this choice? In this section, we present one possible answer to this question via the notion of the distance to calibration [Blasiok et al. 2023a]. We show that the smooth calibration error gives us the best approximation to this ground-truth measure in an information-theoretic sense.

Recall that we defined $\mathcal{D}^*$ to be the joint distribution of $\mathbf{x}, \mathbf{y}^*$, whereas J^* denotes the joint distribution $(p(\mathbf{x}), \mathbf{y}^*)$.

Definition 6.1 *Distance to calibration [Blasiok et al. 2023a].* Given a distribution $\mathcal{D}^*$, define $\mathrm{Cal}(\mathcal{D}^*)$ to be the set of predictors $q : \mathcal{X} \to [0,1]$ such that q is perfectly calibrated under $\mathcal{D}^*$. Define the *true distance to calibration* of the predictor p as

$$\mathrm{dCE}(p, \mathcal{D}^*) = \min_{q \in \mathrm{Cal}(D^*)} d(p, q).$$

This definition formalizes the intuition that a predictor which can be made perfectly calibrated by a small change to its predictions is close to being calibrated. A desirable property that follows immediately from this definition is that the distance to calibration is continuous (unlike ECE). In fact, dCE is Lipschitz continuous: if we chance our predictor p to a different predictor p' that is ε-close to p ($|d(p,p') \leq \varepsilon|$), the distance to calibration can only change by at most ε ($|\mathrm{dCE}(p, \mathcal{D}^*) - \mathrm{dCE}(p', \mathcal{D}^*)| \leq \varepsilon$). This continuity property can be easily proved using the triangle inequality for the metric d.

Despite its intuitiveness and continuity, dCE differs from the other notions of calibration we have seen so far in a crucial way: it depends on the feature space $\mathcal{X}$ (at least, syntactically). This dependence comes about because both the set $\mathrm{Cal}(\mathcal{D}^*)$ of perfectly calibrated predictors and the distance metric d depend on $\mathcal{X}$. The definition of dCE does not give any hints about how one might go about

computing or approximating it.

It is natural to ask to what extent dCE actually depends on the space $\mathcal{X}$, and if it can be approximated by a calibration measure which is independent of $\mathcal{X}$. This leads us to two new definitions.

Definition 6.2 *[Blasiok et al. 2023a].* The *upper distance to calibration* $\overline{\mathrm{dCE}}(J^*)$ is the maximum of $\mathrm{dCE}(p', \mathcal{D}')$ over all spaces $\mathcal{X}'$, distributions $\mathcal{D}'$ on $\mathcal{X}' \times \{0,1\}$ and predictors $p' : \mathcal{X}' \to [0,1]$ such that the distribution $J' = (p'(\mathbf{x}'), \mathbf{y}')$ is identical to the distribution $J^* = (\mathbf{p}(\mathbf{x}), \mathbf{y}^*)$. The *lower distance to calibration* $\underline{\mathrm{dCE}}$ is defined analogously, replacing the maximum by minimum.

By their definition, both $\underline{\mathrm{dCE}}$ and $\overline{\mathrm{dCE}}$ achieve the goal of only depending on J^* and not $\mathcal{D}^*$. It also follows that

$$\underline{\mathrm{dCE}}(J^*) \le \mathrm{dCE}(p, \mathcal{D}^*) \le \overline{\mathrm{dCE}}(J^*).$$

This leads to two questions:

(1) The definitions of $\underline{\mathrm{dCE}}$ and $\overline{\mathrm{dCE}}$ seem rather cumbersome at first, since they involve optimizing over a possibly infinite collection of feature spaces and predictors. Are there more tractable characterizations of these notions, ideally ones that will let us estimate them efficiently?

(2) How far apart are $\underline{\mathrm{dCE}}$ and $\overline{\mathrm{dCE}}$? An ideal situation would be that they are always equal, or at most a constant factor apart. If so, either of them could serve as a good approximation for dCE, assuming we find efficient ways to compute them.

In the following subsection, we will show that the largest gap between the upper and lower distances is quadratic ($\overline{\mathrm{dCE}}(J^*) \le 4\sqrt{\underline{\mathrm{dCE}}(J^*)}$), and that the smooth calibration error gives a constant-factor approximation to the lower distance to calibration. Together, these results let us efficiently approximate the distance to calibration using smooth calibration error, as in the work of [Hu et al. 2024].

6.1 Characterizing and Relating the Upper and Lower Distances to Calibration

In this subsection, we answer the two questions above. Specifically, we give simple characterizations for the upper and lower distances in Theorems 6.6 and 6.7. We show that the two distances are at most quadratically apart in Theorem 6.9.

We first give a simpler characterization of the upper distance. We begin with some definitions needed to state the characterization.

Definition 6.3 *Calibrated post-processing.* Define the set $K(J^*)$ to be the set of post-processing functions that, when applied to p, give a perfectly calibrated predictor. Formally, $K(J^*) = \{\kappa : [0,1] \to [0,1]\ s.t.\ (\kappa(p(\mathbf{x})), y^*)$ is perfectly calibrated.$\}$

We observe that the set $K(J^*)$ is non-empty, since the constant predictor which predicts $\mathbb{E}[y^*]$ is calibrated, and this corresponds to the constant function $\kappa^{\mathsf{av}}(v) = \mathbb{E}[\mathbf{y}^*]$ for all v. A more interesting post-processing is $\kappa^{\mathsf{recal}}(v) = \mathbb{E}[\mathbf{y}^* | v]$, and we call the post-processed predictor $\widehat{p}(\mathbf{x}) := \kappa^{\mathsf{recal}}(p(\mathbf{x}))$ the *recalibration* of p: this predictor keeps the same level sets as p, and changes the predictions to be calibrated.

Definition 6.4 *Recalibration.* Fix a distribution $\mathcal{D}$ of $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \{0,1\}$. We define the *recalibration* of a predictor $p : \mathcal{X} \to [0,1]$ to be another predictor, denoted by $\widehat{p} : \mathcal{X} \to [0,1]$, where $\widehat{p}(\mathbf{x}) := \mathbb{E}_{\mathcal{D}}[\mathbf{y}|p(\mathbf{x})]$.

LEMMA 6.5. *It holds that* $\mathrm{ECE}(p, \mathcal{D}^*) = d(p, \widehat{p}) = d(p, \kappa^{\mathsf{recal}} \circ p)$*, where* $\circ$ *denotes function composition.*

In general, the set $K(J^*)$ could be much richer and possibly induce *closer* calibrated predictors. In particular, there often exist post-processings $\kappa \in K(J^*)$ such that $d(p, \kappa \circ p)$ is much smaller than $d(p, \kappa^{\mathsf{recal}} \circ p) = \mathrm{ECE}(p, \mathcal{D}^*)$. For the two point distribution $\mathcal{D}_2$ considered before, we have seen that $\mathrm{ECE}(p, \mathcal{D}_2) = 1/2 - \epsilon$ whereas it follows that $\kappa^{av} = 1/2$ and $d(p, 1/2) = \epsilon$.

[Blasiok et al. 2023a] give the following characterization of the upper distance.

THEOREM 6.6. *[Blasiok et al. 2023a] We have*

$$\overline{\mathrm{dCE}}(J^*) = \min_{\kappa \in K(J^*)} d(p, \kappa \circ p) = \min_{\kappa \in K(J^*)} \mathop{\mathbb{E}}_{\mathbf{x}} |\kappa(p(\mathbf{x})) - p(\mathbf{x})|.$$

This theorem tells us that the upper distance of a given predictor p is exactly its distance to the closest perfectly calibrated predictor that can be obtained by applying a post-processing κ to p.

Let us sketch the proof idea. $K(J^*)$ is the set of relabelings of the level sets of p which result in a calibrated predictor. For any space X', distribution D' and predictor p' where $J' = J^*$, applying the post-processing function $\kappa \in J^*$ results in a perfectly calibrated predictor $\kappa(p')$ on $\mathcal{X}'$. Hence the distance from such predictors is always an upper bound on $\overline{\mathrm{dCE}}$. For the space X'' where each level set is a single point, these are the only calibrated predictors, so the bound is tight.

We now turn to the lower distance. The good news is that the characterization is in terms of a calibration measure that we have encountered previously: the smooth calibration error $\mathrm{smCE}(p, \mathcal{D}^*)$. The proof however is more involved, we refer the reader to [Blasiok et al. 2023a; Blasiok and Nakkiran 2024].

THEOREM 6.7 [BLASIOK ET AL. 2023A]. *We have*

$$\mathrm{smCE}(p, \mathcal{D}^*)/2 \leq \underline{\mathrm{dCE}}(J^*) \leq 2\mathrm{smCE}(p, \mathcal{D}^*)$$

This theorem lets us efficiently approximate the lower distance to calibration, up to a constant factor, by computing the smooth calibration error. An efficient algorithm for computing the smooth calibration error is given by [Hu et al. 2024].

We now address the question of how close the upper and lower distances are. Assume that all we know about the predictor p and distribution $\mathcal{D}^* = (\mathbf{x}, \mathbf{y}^*)$ is the distribution $J^* = (p(\mathbf{x}), \mathbf{y}^*)$. Does this specify $\mathrm{dCE}(p, D^*)$ completely? Or is there still some uncertainty about how far the closest calibrated predictor is, depending on the space $\mathcal{X}$? The answer (perhaps surprisingly) is that there is quadratic uncertainty in the distance, given J^*.

COROLLARY 6.8. *No calibration measure based on* J^* *can distinguish between the cases where* $\mathrm{dCE}(p, \mathcal{D}^*) \geq \eta$ *and* $\mathrm{dCE}(p, \mathcal{D}^*) \leq 2\eta^2$.

We present an example illustrating Corollary 6.8 in Appendix B. Specifically, we construct pairs of predictors and distributions $(p_1, \mathcal{D}_1^*)$ and $(p_2, \mathcal{D}_2^*)$ so that J^* is

identical in both cases, but dCE differs by a quadratic factor. It turns out that this quadratic separation is in fact the worst possible.

THEOREM 6.9 [BLASIOK ET AL. 2023A]. *We have* $\overline{\mathrm{dCE}}(J^*) \leq 4\sqrt{\underline{\mathrm{dCE}}(J^*)}$.

We discuss the proof of this theorem in Appendix C following the original approach of [Blasiok et al. 2023a] via the notion of *interval calibration error*.

Conclusion.

The classic notion of calibration needs to be rethought in order to satisfy requirements like robustness and computational efficiency, motivated by applications to machine learning and decision making. This leads to a rich set of new questions, in terms of what are desirable properties for approximate calibration notions to have and new algorithmic challenges that arise from trying to achieve these properties. This is a broad and active area of research that spans machine learning, decision making and computational complexity. There are several questions that still remain, such as efficient and meaningful notions of calibration for the multiclass setting [Gopalan et al. 2024] and the generative setting [Kalai and Vempala 2024]. We hope to have given the reader a feel for this in the survey, by highlighting the motivating questions, the definitional challenges and the algorithmic issues.

REFERENCES

ARUNACHALESWARAN, E. R., COLLINA, N., ROTH, A., AND SHI, M. 2025. An elementary predictor obtaining $2\sqrt{T}+1$ distance to calibration. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2025, New Orleans, LA, USA, January 12-15, 2025*, Y. Azar and D. Panigrahi, Eds. SIAM, 1366–1370.

BLASIOK, J., GOPALAN, P., HU, L., AND NAKKIRAN, P. 2023a. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023*. ACM, 1727–1740.

BLASIOK, J., GOPALAN, P., HU, L., AND NAKKIRAN, P. 2023b. When does optimizing a proper loss yield calibration? In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. Vol. 36. Curran Associates, Inc., 72071–72095.

BLASIOK, J. AND NAKKIRAN, P. 2024. Smooth ECE: principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

CASACUBERTA, S., DWORK, C., AND VADHAN, S. 2024. Complexity-theoretic implications of multicalibration. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. STOC 2024. Association for Computing Machinery, New York, NY, USA, 1071–1082.

DAGAN, Y., DASKALAKIS, C., FISHELSON, M., GOLOWICH, N., KLEINBERG, R., AND OKOROAFOR, P. 2025. Breaking the t^(2/3) barrier for sequential calibration. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing, STOC 2025, Prague, Czechia, June 23-27, 2025*, M. Koucký and N. Bansal, Eds. ACM, 2007–2018.

DWORK, C., KIM, M. P., REINGOLD, O., ROTHBLUM, G. N., AND YONA, G. 2021. Outcome indistinguishability. In *ACM Symposium on Theory of Computing (STOC'21)*.

DWORK, C., LEE, D., LIN, H., AND TANKALA, P. 2023. From pseudorandomness to multi-group fairness and back. In *Proceedings of Thirty Sixth Conference on Learning Theory*, G. Neu and L. Rosasco, Eds. Proceedings of Machine Learning Research, vol. 195. PMLR, 3566–3614.

FISHELSON, M., GOLOWICH, N., MOHRI, M., AND SCHNEIDER, J. 2025. High-dimensional calibration from swap regret. *arXiv preprint arXiv:2505.21460*.

FOSTER, D. P. AND VOHRA, R. V. 1998. Asymptotic calibration. *Biometrika 85,* 2, 379–390.

GNEITING, T. AND RAFTERY, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102,* 477, 359–378.

Gopalan, P., Hu, L., and Rothblum, G. N. 2024. On computationally efficient multi-class calibration. In *The Thirty Seventh Annual Conference on Learning Theory*. Proceedings of Machine Learning Research, vol. 247. PMLR, 1983–2026.

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., and Wieder, U. 2022. Omnipredictors. In *Innovations in Theoretical Computer Science (ITCS'2022)*.

Gopalan, P., Kim, M. P., Singhal, M., and Zhao, S. 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.

Hartline, J., Wu, Y., and Yang, Y. 2025. Smooth Calibration and Decision Making. In *6th Symposium on Foundations of Responsible Computing (FORC 2025)*, M. Bun, Ed. Leibniz International Proceedings in Informatics (LIPIcs), vol. 329. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 16:1–16:26.

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML*.

Hu, L., Jambulapati, A., Tian, K., and Yang, C. 2024. Testing calibration in nearly-linear time. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Kakade, S. and Foster, D. 2008. Deterministic calibration and Nash equilibrium. *Journal of Computer and System Sciences 74(1)*, 115–130.

Kalai, A. T. and Vempala, S. S. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. STOC 2024. Association for Computing Machinery, New York, NY, USA, 160–171.

Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F., and Reingold, O. 2022. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences 119,* 4.

Kleinberg, B., Leme, R. P., Schneider, J., and Teng, Y. 2023. U-calibration: Forecasting for an unknown agent. In *Proceedings of Thirty Sixth Conference on Learning Theory*, G. Neu and L. Rosasco, Eds. Proceedings of Machine Learning Research, vol. 195. PMLR, 5143–5145.

Kumar, A., Sarawagi, S., and Jain, U. 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. PMLR, 2805–2814.

Li, Y., Hartline, J. D., Shan, L., and Wu, Y. 2022. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. Association for Computing Machinery, New York, NY, USA, 988–989.

McCarthy, J. 1956. Measures of the value of information. *Proceedings of the National Academy of Sciences 42,* 9, 654–655.

Okoroafor, P., Kleinberg, R., and Kim, M. P. 2025. Near-optimal algorithms for omniprediction. *arXiv preprint arXiv:2501.17205*.

Peng, B. 2025. High dimensional online calibration in polynomial time. *arXiv preprint arXiv:2504.09096*.

Qiao, M. and Valiant, G. 2021. Stronger calibration lower bounds via sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2021. Association for Computing Machinery, New York, NY, USA, 456–466.

Qiao, M. and Zheng, L. 2024. On the distance from calibration in sequential prediction. In *Proceedings of Thirty Seventh Conference on Learning Theory*, S. Agrawal and A. Roth, Eds. Proceedings of Machine Learning Research, vol. 247. PMLR, 4307–4357.

Rossellini, R., Soloff, J. A., Barber, R. F., Ren, Z., and Willett, R. 2025. Can a calibration metric be both testable and actionable? *arXiv preprint arXiv:2502.19851*.

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

A. DEFERRED PROOFS

A.1 Proof of Lemma C.2

Proof. Let $w_j = \mathbb{E}[p(\mathbf{x})|p(\mathbf{x}) \in I_j]$, and note that $w_j \in I_j$, a property that will be used shortly. We can write

$$\text{intCE}_B(p, J^*) = \text{width}(B) + \sum_j \Pr[p(\mathbf{x}) \in I_j]|v_j - w_j|. \tag{A.1}$$

We now bound $d(p, q_B)$ as

$$\begin{aligned}
d(p, q_B) &= \mathop{\mathbb{E}}_{\mathcal{D}^*}[|p(\mathbf{x}) - q_B(x)|] \\
&= \sum_{j\in[k]} \Pr[p(\mathbf{x}) \in I_j]\, \mathbb{E}[|p(\mathbf{x}) - v_j||p(\mathbf{x}) \in I_j] \\
&\le \sum_{j\in[k]} \Pr[p(\mathbf{x}) \in I_j]\, (\mathbb{E}[|p(\mathbf{x}) - w_j||p(\mathbf{x}) \in I_j] + |w_j - v_j|) \\
&\le (\sum_{j\in[k]} \Pr[p(\mathbf{x}) \in I_j])\text{width}(B) + \sum_{j\in[k]} \Pr[p(\mathbf{x}) \in I_j]|v_j - w_j| \\
&= \text{intCE}_B(p, J^*) \quad \text{(By equation(A.1))}
\end{aligned}$$

where the penultimate line uses the fact that conditioned on $p(\mathbf{x}) \in I_j$, $|p(\mathbf{x}) - w_j| \le \text{width}(B)$ since both values lie in the interval I_j. □

A.2 Proof of Lemma C.4

Proof. Let β be a width parameter to be chosen later. We consider the bucketing B where the first interval is $[0, b]$ for b picked randomly from the interval $[0, \beta]$. Every subsequent interval has width β (except possibly the last, which might be smaller). Denote the intervals by $I_1, \ldots, I_k$.

For the predictor q, the calibration error term for B is 0 since

$$\text{CE}_B(q) = \sum_{j\in k} |\,\mathbb{E}[\mathbf{1}(q(\mathbf{x}) \in I_j)(\mathbf{y}^* - q(\mathbf{x}))]| \le \int_{v\in[0,1]} \Pr[q(\mathbf{x}) = v]|\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x})|q(\mathbf{x}) = v]| = 0.$$

So we will try the bound the calibration term for p by comparing it to q and arguing that if they are close by, this error is small.

$$\begin{aligned}
\text{CE}_B(p, \mathcal{D}^*) &= \sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - p(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]| \\
&\le \sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]| + \sum_{j\in k} |\,\mathbb{E}[(q(\mathbf{x}) - p(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]|
\end{aligned} \tag{A.2}$$

We bound each of these terms separately. To bound the second term,

$$\sum_{j\in k} |\,\mathbb{E}[(q(\mathbf{x}) - p(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]| = \mathbb{E}[|q(x) - p(x)|] \le \delta \tag{A.3}$$

For the first term, we have

$$\begin{aligned}\sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]| &\leq \sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))\mathbb{I}(q(\mathbf{x}) \in I_j)]| + \\ &\quad |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))(\mathbb{I}(p(\mathbf{x}) \in I_j) - \mathbb{I}(q(\mathbf{x}) \in I_j)] \\ &\leq \sum_j |\mathbb{I}(p(\mathbf{x}) \in I_j) - \mathbb{I}(q(\mathbf{x}) \in I_j)|\end{aligned}$$

where we use $\mathrm{CE}_B(q, \mathcal{D}^*) = 0$ and $|\mathbf{y}^* - q(\mathbf{x})| \leq 1$. The RHS is 0 if $p(\mathbf{x})$ and $q(\mathbf{x})$ land in the same bucket, else it is 2. $p(\mathbf{x})$ and $q(\mathbf{x})$ land in different buckets if there is a bucket boundary between them, which happens with probability bounded by $|p(\mathbf{x}) - q(\mathbf{x})|/\beta$ over the random choice of b. Hence we can bound

$$\sum_{j\in k} |\,\mathbb{E}[(\mathbf{y}^* - q(\mathbf{x}))\mathbb{I}(p(\mathbf{x}) \in I_j)]| \leq \frac{2\,\mathbb{E}[|p(\mathbf{x}) - q(\mathbf{x})|]}{\beta} = \frac{2\delta}{\beta}. \tag{A.4}$$

Plugging Equations (A.3) and (A.4) back into Equation (A.2) and choosing $\beta = \sqrt{2\delta}$,

$$\begin{aligned}\mathrm{CE}_B(p, \mathcal{D}^*) &\leq \delta + 2\delta/\beta. \\ \mathrm{intCE}_B(p, \mathcal{D}^*) &\leq \mathrm{CE}_B(p, \mathcal{D}^*) + \mathrm{width}(B) \leq \beta + \delta + 2\sqrt{\delta} \leq 4\sqrt{\delta}.\end{aligned}$$

□

A.3 Proof of Lemma 4.4

Proof. The lemma can be proved by considering the level sets $\mathcal{X}_v := \{x \in \mathcal{X} : p(x) = v\}$ for $v \in [0, 1]$. Within each level set, p is a constant function, and the functions $\sigma(p(x))$ formed by all choices of $\sigma : [0, 1] \to A$ are all the constant functions on this level set taking value in A. Moreover, for any level set $\mathcal{X}_v$, the conditional distribution of $\mathbf{y}$ given $\mathbf{x} \in \mathcal{X}_v$ is the Bernoulli distribution with parameter $\widehat{p}(\mathbf{x})$, where $\widehat{p}(x)$ is also a constant function for $x \in \mathcal{X}_v$. Decomposing (4.4) by the level sets, the lemma follows from the definition of the best-response function $\sigma_{\mathcal{T}}$ in (4.1). □

A.4 Proof of Lemma 4.7

Proof. By Theorem 4.6,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x}), \mathbf{y}))] &= \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(p(\mathbf{x})) + \nabla\varphi(p(\mathbf{x})) \cdot (\mathbf{y} - p(\mathbf{x}))] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(p(\mathbf{x})) + \nabla\varphi(p(\mathbf{x})) \cdot (\widehat{p}(\mathbf{x}) - p(\mathbf{x}))] \\ &\qquad\qquad \text{(because } \mathbb{E}[\mathbf{y}|p(\mathbf{x})] = \widehat{p}(\mathbf{x})\text{)} \\ &= \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x})) - \varphi(p(\mathbf{x})) - \nabla\varphi(p(\mathbf{x})) \cdot (\widehat{p}(\mathbf{x}) - p(\mathbf{x}))] \\ &= \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[D_\varphi(\widehat{p}(\mathbf{x}) \| p(\mathbf{x}))].\end{aligned}$$

This proves Equation (4.7). Similarly,

$$\mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_{\mathcal{T}}(\widehat{p}(\mathbf{x}), \mathbf{y}))] = \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[D_\varphi(\widehat{p}(\mathbf{x}) \| \widehat{p}(\mathbf{x}))] = \mathop{\mathbb{E}}_{\mathcal{D}}[\varphi(\widehat{p}(\mathbf{x}))].$$

Taking the difference between the two equations above, we have

$$\mathrm{CFDL}_{\mathcal{T}}(p,\mathcal{D}) = \mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_{\mathcal{T}}(\widehat{p}(\mathbf{x}),\mathbf{y}))] - \mathop{\mathbb{E}}_{\mathcal{D}}[u(\sigma_{\mathcal{T}}(p(\mathbf{x}),\mathbf{y}))] = \mathop{\mathbb{E}}_{\mathcal{D}}[D_{\varphi}(\widehat{p}(\mathbf{x})\|p(\mathbf{x}))].$$

This proves Equation (4.8). □

A.5 Proof of Theorem 4.9

Proof. Let $\widehat{p}$ be the recalibration of p (Definition 6.4). By the definitions of J^* and J^p, for any $x \in \mathcal{X}$, we have

$$\mu_{J^*}(p(x)) = \widehat{p}(x), \tag{A.5}$$

$$\mu_{J^p}(p(x)) = p(x). \tag{A.6}$$

Let M denote the marginal distribution of $p(\mathbf{x})$ where $(\mathbf{x},\mathbf{y}^*) \sim \mathcal{D}$. By Lemma 4.7,

$$\begin{aligned}\mathrm{CFDL}_{\mathcal{T}}(p,\mathcal{D}) &= \mathop{\mathbb{E}}_{\mathcal{D}}[D_{\varphi}(\widehat{p}(\mathbf{x})\|p(\mathbf{x}))] \\ &= \mathop{\mathbb{E}}_{\mathbf{v}\sim M}[D_{\varphi}(\mu_{J^*}(\mathbf{v})\|\mu_{J^p}(\mathbf{v}))] \\ &= D_{\varphi}(J^*\|J^p).\end{aligned}$$

□

A.6 V-shaped Divergences

In this subsection, we discuss a fundamental result about Bregman divergences (Theorem A.1) that will be used to prove Theorem 4.3.

CDL focuses on decision tasks $\mathcal{T} = (A,u)$ with $[0,1]$-bounded payoff functions $u : A \to [0,1]$. For such tasks, the corresponding convex function φ from Theorem 4.6 must have bounded subgradients:

$$\nabla\varphi(v) = u(\sigma_{\mathcal{T}}(v),1) - u(\sigma_{\mathcal{T}}(v),0) \in [-1,1] \quad \text{for every } v \in [0,1]. \tag{A.7}$$

While the convex functions φ with bounded subgradients $\nabla\varphi(v) \in [-1,1]$ form a large family, a fundamental result by [Li et al. 2022], which we include as Theorem A.1 below, shows that the divergences D_{φ} defined by this family can be captured by extremely simple functions φ that are termed *V-shaped* functions. Specifically, for each $v^* \in [0,1]$, a V-shaped function φ_{v^*} is defined as follows:

$$\varphi_{v^*}(v) = |v - v^*| \quad \text{for every } v \in [0,1].$$

The Bregman divergence $D_{\varphi_{v^*}}$ is correspondingly termed a *V-shaped* divergence, and it can be easily computed as follows: for $v_1, v_2 \in [0,1]$, we have

$$D_{\varphi_{v^*}}(v_1\|v_2) = \begin{cases} 2|v_1 - v^*| \le 2|v_1 - v_2|, & \text{if } v^* \in (v_1,v_2] \text{ or if } v^* \in (v_2,v_1]; \\ 0, & \text{otherwise.} \end{cases} \tag{A.8}$$

The following theorem gives an upper bound on the expected divergence D_{φ} for a general φ with bounded subgradient in terms of V-shaped divergences $D_{\varphi_{v^*}}$.

Theorem A.1 [Li et al. 2022]. *Let $\varphi : [0,1] \to \mathbb{R}$ be a convex function whose subgradient is bounded: $\nabla\varphi(v) \in [-1,1]$ for every $v \in [0,1]$. Then for any distribution Π of $(v_1,v_2) \in [0,1]$,*

$$\mathop{\mathbb{E}}_{(v_1,v_2)\sim\Pi} D_{\varphi}(v_1,v_2) \le \sup_{v^*\in[0,1]} \mathop{\mathbb{E}}_{(v_1,v_2)\sim\Pi} D_{\varphi_{v^*}}(v_1,v_2).$$

A.7 Relationship to ECE

We prove Theorem 4.3, which demonstrates the quadratic relationship between the CDL and the ECE. The first and last inequalities in (4.3) follow immediately from Jensen's inequality. The second inequality can be proved using the decision task $\mathcal{T}_2$ from Section 4.2. Specifically, the payoff function of $\mathcal{T}_2$ has its range bounded in $[0,1]$, so by the definition of CDL and Equation (4.6),

$$\mathrm{CDL}(p,\mathcal{D}) \geq \mathrm{CFDL}_{\mathcal{T}_2}(p,\mathcal{D}) = \mathrm{ECE}_2(p,\mathcal{D})^2.$$

Now we prove the third inequality in (4.3). By Lemma 4.7 and Theorem A.1, for any decision task $\mathcal{T}$ with $[0,1]$ bounded payoffs,

$$\begin{aligned}\mathrm{CFDL}_{\mathcal{T}}(p,\mathcal{D}) = \mathop{\mathbb{E}}_{\mathcal{D}}[D_{\varphi}(\widehat{p}(\mathbf{x})\|p(\mathbf{x}))] &\leq \sup_{v^*\in[0,1]} \mathop{\mathbb{E}}_{\mathcal{D}}[D_{\varphi_{v^*}}(\widehat{p}(\mathbf{x})\|p(\mathbf{x}))] \\ &\leq 2\mathop{\mathbb{E}}_{\mathcal{D}}|\widehat{p}(\mathbf{x}) - p(\mathbf{x})| = \mathrm{ECE}(p,\mathcal{D}). \quad \text{(by (A.8))}\end{aligned}$$

This proves $\mathrm{CDL}(p,\mathcal{D}) \leq 2\mathrm{ECE}(p,\mathcal{D})$, as desired.

A.8 Tight examples between CDL and ECE

We prove the correctness of the two examples $(p_1,\mathcal{D}_1),(p_2,\mathcal{D}_2)$ we mentioned after Theorem 4.3 that shows the tightness of Theorem 4.3.

In the first example, we have $p_1(x) = 1/2$ and $\widehat{p}_1(x) = 1/2+\varepsilon$ for any x, so it is clear that $\mathrm{ECE}_2(p_1,\mathcal{D}_1) = \varepsilon$. To prove $\mathrm{CDL}(p_1,\mathcal{D}_1) \geq 2\varepsilon$, consider the task $\mathcal{T}_1 = (A,u)$ with two actions: $A = \{0,1\}$. The payoff function u is defined such that $u(a,y) = 1$ if $a = y$, and $u(a,y) = 0$ otherwise. The best-response function is $\sigma_{\mathcal{T}}(v) = 0$ if $v \leq 1/2$, and $\sigma_T(v) = 1$ otherwise. We have

$$\begin{aligned}\mathbb{E}[u(\sigma_{\mathcal{T}_1}(p_1(\mathbf{x})),\mathbf{y})] &= \mathbb{E}[u(0,\mathbf{y})] = \Pr[\mathbf{y}=0] = \frac{1}{2} - \varepsilon, \\ \mathbb{E}[u(\sigma_{\mathcal{T}_1}(\widehat{p}_1(\mathbf{x})),\mathbf{y})] &= \mathbb{E}[u(1,\mathbf{y})] = \Pr[\mathbf{y}=1] = \frac{1}{2} + \varepsilon.\end{aligned}$$

Taking the difference between the two expected payoffs, we get $\mathrm{CDL}(p_1,\mathcal{D}_1) \geq \mathrm{CFDL}_{\mathcal{T}_1}(p_1,\mathcal{D}_1) = 2\varepsilon$.

In the second example, we have $p_2(x) = x$ and $\widehat{p}_2(x) = x - \varepsilon$, so it is clear that $\mathrm{ECE}(p_2,\mathcal{D}_2) = \varepsilon$. Now we prove

$$\mathrm{CDL}(p_2,\mathcal{D}_2) \leq \frac{\varepsilon^2}{1-\varepsilon} = \varepsilon^2 + O(\varepsilon^3). \tag{A.9}$$

Consider any decision task $\mathcal{T} = (A,u)$ with a $[0,1]$-bounded payoff function u :

$A \to [0,1]$. By Lemma 4.7 and Theorem A.1,

$$\begin{aligned}
\mathrm{CFDL}_{\mathcal{T}}(p_2, \mathcal{D}_2) &= \mathop{\mathbb{E}}_{\mathcal{D}_2}[D_\varphi(\widehat{p}_2(\mathbf{x}) \| p_2(\mathbf{x}))] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}_2}[D_\varphi(\mathbf{x} - \varepsilon \| \mathbf{x})] \\
&\leq \sup_{v^* \in [0,1]} \mathop{\mathbb{E}}_{\mathcal{D}_2}[D_{\varphi_{v^*}}(\mathbf{x} - \varepsilon \| \mathbf{x})] \\
&= \sup_{v^* \in [0,1]} \Pr_{\mathcal{D}_2}\left[v^* - (\mathbf{x} - \varepsilon) \middle| v^* \in (\mathbf{x} - \varepsilon, \mathbf{x}]\right] \qquad \text{(by (A.8))} \\
&= \sup_{v^* \in [0,1]} \int_0^1 (v^* - (x - \varepsilon)) \mathbb{I}(v^* \in (x - \varepsilon, x]) \mathrm{d}x \\
&\leq \sup_{v^* \in [0,1]} \int_{v^*}^{v^* + \varepsilon} (v^* - (x - \varepsilon)) \mathrm{d}x \\
&= 2\varepsilon^2. \qquad \text{(A.10)}
\end{aligned}$$

Since this upper bound on the CFDL holds for any decision task $\mathcal{T}$ with a $[0,1]$-bounded payoff function, it implies (A.9), as desired.

B. THE INHERENT UNCERTAINTY IN DISTANCE TO CALIBRATION

Assume that all we know about the predictor p and distribution $\mathcal{D}^* = (\mathbf{x}, \mathbf{y}^*)$ is the distribution $J^* = (p(\mathbf{x}), \mathbf{y}^*)$. Does this specify $\mathrm{dCE}(p, D^*)$ completely? Or is there still some uncertainty on how far the closest calibrated predictor is, depending on the space $\mathcal{X}$?

We present a simple example showing that there is indeed some uncertainty. Take ϵ to be any value in $(0, 1/2)$, and let $\delta = \epsilon/(1 - 2\epsilon)$. The distribution J^* is easy to describe: $p(\mathbf{x})$ takes the values $1/2 + \delta$ and $1/2 - \delta$ each with probability $1/2$, and conditioned on each value of $p(\mathbf{x})$, $\mathbf{y}^*$ is uniformly distributed in $\{0, 1\}$.

Note that any such p is not perfectly calibrated. But it is δ far from the constant $1/2$ predictor, which is perfectly calibrated. It is easy to construct a space where this is indeed the closest calibrated predictor, so that $\mathrm{dCE}(p, \mathcal{D}^*) = \delta$.

What is perhaps less obvious is there exist spaces and predictors realizing J^* where the true distance to calibration is much smaller. We describe one such construction. Let $\mathcal{X} = \{00, 01, 10, 11\}$. Cosndier the distribution $\mathcal{D}^*$ on pairs $(\mathbf{x}, \mathbf{y}^*) \in \mathcal{X} \times \{0, 1\}$, and predictors $p_1, p_2 : \mathcal{X} \to [0, 1]$ given below:

x	$\Pr_{\mathcal{D}^*}[\mathbf{x} = x]$	$\mathbb{E}_{\mathcal{D}^*}[\mathbf{y}^* \mid \mathbf{x} = x]$	$p_1(x)$	$p_2(x)$
00	$\frac{1}{2} - \epsilon$	$\frac{1}{2} - \delta$	$\frac{1}{2} - \delta$	$\frac{1}{2} - \delta$
01	ϵ	1	$\frac{1}{2} - \delta$	$\frac{1}{2}$
10	ϵ	0	$\frac{1}{2} + \delta$	$\frac{1}{2}$
11	$\frac{1}{2} - \epsilon$	$\frac{1}{2} + \delta$	$\frac{1}{2} + \delta$	$\frac{1}{2} + \delta$

The predictor p_1 is not perfectly calibrated, indeed we have chosen δ such that the joint distribution of $(p_1(\mathbf{x}), \mathbf{y}^*)$ is exactly J^*: conditioned on either prediction value in $\{1/2 \pm \delta\}$, the bit $\mathbf{y}^*$ is uniformly random. In contrast, the predictor p_2 is easily seen to be calibrated.

Observe that p_1 and p_2 agree on 00 and 11. They disagree by δ on 01 and 10, which each have ϵ probability under $\mathcal{D}^*$, so $d(p_1, p_2) = 2\epsilon\delta = \Theta(\epsilon^2)$. This establishes the difficulty of pinning down the true distance to calibration within a quadratic factor.

C. RELATING UPPER AND LOWER DISTANCES TO CALIBRATION

In this section, we prove Theorem 6.9 showing that the upper and lower distance to calibration can be at most quadratically far apart. This shows that the simple example in Appendix B is nearly tight. We follow the proof strategy of [Blasiok et al. 2023a] using the notion of *interval calibration error*.

C.1 Interval Calibration Error

Definition C.1 *Interval Calibration Error [Blasiok et al. 2023a].* A interval partition B is a partition of the interval $[0, 1]$ into disjoint intervals $I_1, \ldots, I_k$. We let the width of the partition width(B) be the length of longest interval. Given a predictor p, we define its calibration error and interval calibration error for B respectively as

$$\mathrm{CE}_B(p, \mathcal{D}^*) = \sum_{j \in [k]} |\, \mathbb{E}[(\mathbf{y}^* - p(\mathbf{x}))\mathbf{1}(p(\mathbf{x}) \in I_j)]|$$
$$\mathrm{intCE}_B(p, \mathcal{D}^*) = \mathrm{CE}_B(p, J^*) + \mathrm{width}(B).$$

The *interval calibration error* minimizes over all interval partitions B:

$$\mathrm{intCE}(p, \mathcal{D}^*) = \min_B \mathrm{intCE}_B(p, \mathcal{D}^*).$$

The definition of intCE_B involves two terms that represent a tradeoff: the calibration error term, and the width term that penalizes partitions which use large width intervals. Intuitively, as the intervals grow larger it is easier to reduce calibration error, since we are allowed to cancel out the point-wise errors $\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})] - p(\mathbf{x})$ over larger intervals; but the width penalty also grows larger. At one extreme, we can think of the width 0 case as corresponding to the ECE. At the other extreme, by taking the single interval $[0, 1]$, we pay $\mathbb{E}[\mathbf{y}^* - p(\mathbf{x})]$ which is 0 if the expectations of $\mathbf{y}^*$ and $p(\mathbf{x})$ are equal; a very weak calibration guarantee. But now the width penalty is 1.

Formal justification for the definition comes from the following observation. The canonical predictor q_B for an interval partition B and a distribution $\mathcal{D}^*$ is the predictor where for all $x \in I_j$, the q_B predicts $v_j = \mathbb{E}[y^*|p(\mathbf{x}) \in I_j|$. It is easy to see that q_B is perfectly calibrated for $\mathcal{D}^*$.

Lemma C.2. *The canonical predictor q_B for $B, \mathcal{D}^*$ satisfies $d(p, q_B) \leq \mathrm{intCE}_B(p, \mathcal{D}^*)$.*

This leads to the following upper bound:

Theorem C.3. *[Blasiok et al. 2023a] We have $\overline{\mathrm{dCE}}(p, \mathcal{D}^*) \leq \mathrm{intCE}(p, \mathcal{D}^*)$.*

To prove Theorem C.3 we observe that the canonical predictor q_B can be viewed as a post-processing of the predictor p, since we can write $q_B(x) = \kappa(p(x))$ where $\kappa(t) = v_j$ for $t \in I_j$. Thus by Lemma C.2,

$$\overline{\mathrm{dCE}}(p, \mathcal{D}^*) \leq d(p, q_B) \leq \mathrm{intCE}_B(p, \mathcal{D}^*).$$

Minimizing over all B completes the proof.

The reader might wonder, why define yet another calibration measure? The answer is two-fold:

—Interval calibration error gives a simple yet powerful upper bound on the upper distance to calibration. In the next subsection, this allows us to relate the upper and lower distance to calibration, showing that they are never more than quadratically far apart. This is formally proved in Theorem 6.9, showing the gap example in Corollary 6.8 is the worst possible (up to constants).

—It presents a rigorous alternative to heuristic measures like bucketed ECE: regularize the calibration error by adding the max bucket width. This allows for meaningful comparison of calibration scores obtained using different number or other choice of buckets, rather than leaving the number of buckets as a hyperparameter.

C.2 Proof of Theorem 6.9

Let us pick $\mathcal{X}, \mathcal{D}^*, p$ to be the space, distribution and predictor respectively that achieve the lower distance to calibration for J^*. So there exists a perfectly calibrated predictor $q : \mathcal{X} \to [0,1]$ such that $d(p,q) = \underline{\mathrm{dCE}}(p, \mathcal{D}^*) = \delta$. We wish to infer the existence of a bucketing B so that $\mathrm{intCE}_B(p, \mathcal{D}^*)$ is small. By Theorem C.3, this will imply that the upper distance is bounded. Corollary 6.8 tells us that we cannot hope for an upper bound better than $\sqrt{\delta}/2$. It turns out that this is not far from the best possible (see Section A.2 for proof):

LEMMA C.4. *There exists a bucketing B such that* $\mathrm{intCE}_B(p, \mathcal{D}^*) \le 4\sqrt{\delta}$.

Combining this lemma with Corollary 6.8, we have completed the proof:

$$\overline{\mathrm{dCE}}(p, \mathcal{D}^*) \le \mathrm{intCE}(p, \mathcal{D}^*) \le \mathrm{intCE}_B(p, \mathcal{D}^*) \le 4\sqrt{\delta} = 4\sqrt{\underline{\mathrm{dCE}}(p, \mathcal{D}^*)}.$$