Can a calibration metric be both testable and actionable?

Raphael Rossellini RAPHAELR@UCHICAGO.EDU
Department of Statistics, University of Chicago

Jake A. Soloff SOLOFF@UCHICAGO.EDU
Department of Statistics, University of Chicago

Rina Foygel Barber RINA@UCHICAGO.EDU
Department of Statistics, University of Chicago

Zhimei Ren ZREN@WHARTON.UPENN.EDU
Department of Statistics and Data Science, the Wharton School, University of Pennsylvania

Rebecca Willett WILLETT@UCHICAGO.EDU
Department of Statistics and Computer Science, University of Chicago

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Forecast probabilities often serve as critical inputs for binary decision making. In such settings, calibration—ensuring forecasted probabilities match empirical frequencies—is essential. Although the common notion of Expected Calibration Error (ECE) provides actionable insights for decision making, it is not testable: it cannot be empirically estimated in many practical cases. Conversely, the recently proposed Distance from Calibration (dCE) is testable, but it is not actionable since it lacks decision-theoretic guarantees needed for high-stakes applications. To resolve this question, we consider Cutoff Calibration Error, a calibration measure that bridges this gap by assessing calibration over intervals of forecasted probabilities. We show that Cutoff Calibration Error is both testable and actionable, and we examine its implications for popular post-hoc calibration methods, such as isotonic regression and Platt scaling.

Keywords: Calibration, Decision Theory, Distribution-free

1. Introduction

To what extent should a decision-maker trust probability forecasts? For example, airlines may need to decide whether to cancel flights based on a weather model's predicted probability of severe weather. The reliability of such decisions fundamentally depends on whether these probabilities are *calibrated*—that is, whether an 80% forecast actually corresponds to an 80% chance of the event occurring. Formally, a forecast model $f : \mathcal{X} \to [0, 1]$ is perfectly calibrated if

$$\mathbb{E}[Y \mid f(X)] = f(X) \ \text{ almost surely.} \tag{1}$$

However, many modern forecasting models, including neural networks, are known to produce miscalibrated probabilities (Guo et al., 2017).

A challenge lies in quantifying how close the model f comes to the ideal of perfect calibration (1). There are two main desiderata for any measure of approximate calibration $\Delta(f)$. First, it should be *testable*, in the sense that we should be able to reliably estimate the metric from finite data. In particular, we need to be able to test whether f is approximately calibrated, $\Delta(f) \approx 0$. Second,

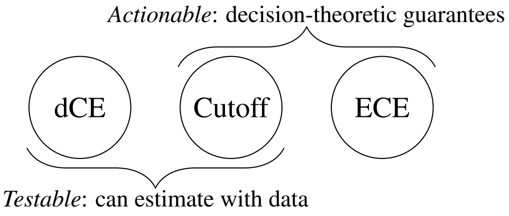
Actionable: decision-theoretic guarantees
dCE
Cutoff
ECE
Testable: can estimate with data

Figure 1: Comparing Cutoff Calibration Error to ECE and dCE.

any calibration metric should be *actionable*—if $\Delta(f)$ is small, this should imply meaningful guarantees for downstream decision-making. For example, if a medical diagnostic system's probabilities have low calibration error, doctors should be confident that acting on its risk predictions according to clinical guidelines will lead to good patient outcomes. Unfortunately, as we will see, these two requirements are in tension: natural measures of calibration that provide strong decision-theoretic guarantees often turn out to be statistically intractable to estimate, and vice versa. We discuss in more detail these desiderata of actionability and testability in Section 3 and Section 4, respectively.

A popular calibration metric is the Expected Calibration Error (ECE):

$$\Delta_{\mathrm{ECE}}(f) := \mathbb{E}\left[\Big|\mathbb{E}\left[Y|f(X)\right] - f(X)\Big|\right].$$

Recent work has observed that ECE is actionable for binary decision problems (Kleinberg et al., 2023; Hu and Wu, 2024). However, ECE has severe limitations around its testability. Standard approaches to estimating $\Delta_{\mathrm{ECE}}(f)$ are known to be very unstable (Kumar et al., 2019; Nixon et al., 2019) and biased (Roelofs et al., 2022). These issues are ultimately not a technical limitation of existing approaches, but rather a consequence fundamental hardness results for estimating and constructing upper confidence bounds for $\Delta_{\mathrm{ECE}}(f)$ (Gupta et al., 2020).

In part to address the limitations of ECE, Błasiok et al. (2023) propose Distance from Calibration (dCE), which measures how far f is from the nearest perfectly calibrated model:

$$\Delta_{\mathrm{dCE}}(f) := \inf_{\substack{g:\mathcal{X}\to[0,1] \\ \mathbb{E}[Y|g(X)]=g(X)}} \mathbb{E}\left[|g(X) - f(X)|\right].$$

dCE is an easier notion of calibration to satisfy, as compared to ECE: indeed, for any f, $\Delta_{\mathrm{dCE}}(f) \leq \Delta_{\mathrm{ECE}}(f)$ (Błasiok et al., 2023, Lemma 4.7). Moreover, dCE solves the issue of testability, since unlike ECE, bounds on it can be estimated from data. However, this statistical advantage comes at a cost—dCE fails to provide the decision-theoretic assurances that made ECE valuable for applications. We show that models with small dCE can still lead to poor decisions.

To bridge the gap between testable and actionable calibration measures, we consider the *Cutoff Calibration Error*, which we can view as an intermediate definition lying between dCE and ECE. This measure assesses calibration over ranges of predicted probabilities rather than at specific values. We prove that Cutoff Calibration Error can be efficiently estimated from data at the parametric rate, while still providing guarantees about decision quality when using the calibrated forecasts. The key insight is that most practical decisions depend on whether a probability exceeds some threshold, making interval-based calibration particularly relevant.

Our key contributions may be summarized as follows.

Popular existing measures face limitations. Building on the results of Gupta et al. (2020), we establish that testing ECE is fundamentally hard by establishing that any calibration method with asymptotically vanishing ECE must have small effective support size. On the other hand, we demonstrate through counterexamples that dCE cannot provide the same decision-theoretic guarantees as ECE. In other words, we will see that ECE is not testable, while dCE is not actionable.

Cutoff Calibration Error is testable and actionable: We define Cutoff Calibration Error and demonstrate that it combines the strongest benefits of ECE and dCE. First, controlling Cutoff Calibration Error implies decision-theoretic guarantees in the binary decision setting under an intuitive monotonicity constraint. Second, like dCE, Cutoff Calibration Error can be efficiently estimated from data using a plug-in approach that's closely related to a proposal of Arrieta-Ibarra et al. (2022). We study relationships between Cutoff Calibration Error, ECE and dCE in detail.[1]

Implications for Platt scaling and isotonic regression: We study popular post-hoc calibration methods of forecast probabilities in a distribution-free context. We provide a finite-sample bound on Cutoff Calibration Error for isotonic regression. By contrast, we exhibit a counterexample for which Platt Scaling does not control Cutoff Calibration Error even asymptotically.

2. Our proposal: Cutoff calibration

Consider a medical diagnostic system that models probabilities of various conditions. While existing calibration measures like ECE focus on the accuracy of specific probability values (e.g., historical outcomes when the model output was exactly 73%), clinicians typically work with standardized risk thresholds—low, medium, and high risk categories that map to different intervention protocols. A doctor rarely needs to distinguish between a 73% and 74% risk, but they need to know whether a patient falls above or below key decision thresholds; perhaps the 80% mark may prompt immediate attention.

This observation—that real-world decisions often depend on probability ranges rather than exact values—motivates us to reframe approximate calibration. Instead of assessing calibration separately at each possible value of $f(X)$, we evaluate calibration over intervals that correspond to potential decision thresholds:

Definition 1 ***Cutoff Calibration Error*** *is defined as follows:*

$$\Delta_{\text{Cutoff}}(f) := \sup_{\text{intervals } I} \left| \mathbb{E}[(Y - f(X))\mathbf{1}_{\{f(X)\in I\}}] \right|.$$

This definition will be easier to work with than ECE, since we do not have to deal directly with the conditional expectation $\mathbb{E}[Y \mid f(X)]$. For a given interval I, the error $|\mathbb{E}[(Y - f(X))\mathbf{1}_{\{f(X)\in I\}}]|$ may be small either because the forecast is unlikely to take values in I or because f is well-calibrated on I. Concretely, if $\Delta_{\text{Cutoff}}(f) \leq \delta$, then for any interval I,

$$\left| \mathbb{E}\left[\, \mathbb{E}[Y \mid f(X)] - f(X) \,\middle|\, f(X) \in I \,\right] \right| \leq \frac{\delta}{\mathbb{P}\{f(X) \in I\}}.$$

[1] After submission of this paper, we learned of two concurrent works (Okoroafor et al., 2025; Qiao and Zhao, 2025) that address similar questions. Each proposes related definitions of calibration, and each considers the benefits of these definitions from the perspectives of a decision-theoretic framework, albeit in distinct ways; we will comment on these similarities and distinctions as appropriate when presenting our results below. In addition, Okoroafor et al. (2025) also considers testability, although our paper focuses more on how testability hardness results motivate the construction of Cutoff Calibration Error.

Since the difficulty of the calibration problem adapts to the probability on each interval, Cutoff Calibration Error is closely related to weighted calibration error (Gopalan et al., 2022). We further explore this connection below.

By searching over intervals $I \subseteq [0,1]$, this definition will be relevant for the binary decision framework, where decision rules take the form $\mathbf{1}_{\{f(X)\geq\tau\}}$, corresponding to $I = [\tau, 1]$. Taking the supremum over *all* intervals allows downstream decision-makers to apply the threshold most relevant to their loss function.

How does Cutoff Calibration Error relate to the existing calibration metrics, ECE and dCE? We establish a precise relationship through the following chain of inequalities.

Proposition 2 *For any* $f : [0,1] \to [0,1]$,

$$\Delta_{\mathrm{dCE}}(f)^2/9 \leq \Delta_{\mathrm{Cutoff}}(f) \leq \Delta_{\mathrm{ECE}}(f).$$

In other words, ECE is a stronger notion of calibration than Cutoff Calibration Error, and Cutoff Calibration Error is a stronger notion of calibration than dCE. These comparisons are strict: as we will see in examples throughout the paper (see also Section A.4), it is possible to construct a function f for which $\Delta_{\mathrm{dCE}}(f)$ is arbitrarily close to zero while $\Delta_{\mathrm{Cutoff}}(f)$ stays bounded away from zero; or similarly, for which $\Delta_{\mathrm{Cutoff}}(f)$ is arbitrarily close to zero while $\Delta_{\mathrm{ECE}}(f)$ stays bounded away from zero. In the next section, we connect each of these metrics to different kinds of weighted calibration error, which further clarifies their relationship by making the definitions more immediately comparable.

2.1. Connections to weighted calibration error

Weighted calibration error refers to a general family of calibration metrics introduced by Gopalan et al. (2022). The definition is closely related to multicalibration (Hébert-Johnson et al., 2018) and multiaccuracy (Kim et al., 2019) in the fairness literature, where the main distinction is that we only weight (or group) different values of X via the predicted probability $f(X)$.

Definition 3 *Let* $\mathcal{C} \subseteq [-1,1]^{f(\mathcal{X})}$. *Then, the* ***weighted calibration error*** *of* f *with respect to the class* $\mathcal{C}$ *is defined as the following:*

$$\Delta_{\mathrm{wCE}}(f;\mathcal{C}) := \sup_{w\in\mathcal{C}} \mathbb{E}[w(f(X))(Y - f(X))].$$

The various calibration metrics we have considered thus far—dCE, ECE, and Cutoff Calibration Error—all have close connections to weighted calibration error for different choices of the function class $\mathcal{C}$. For example, $\Delta_{\mathrm{dCE}}(f)$ can be thought of as a 1-Wasserstein distance between f and the class of all perfectly calibrated functions. Błasiok et al. (2023) thus note that, by Kantorovich-Rubinstein duality, $\Delta_{\mathrm{dCE}}(f)$ is polynomially equivalent to $\Delta_{\mathrm{wCE}}(f;\mathcal{L}_1)$ where $\mathcal{L}_1$ is the set of all 1-Lipschitz weighting functions $w : [0,1] \to [-1,1]$:

$$\Delta_{\mathrm{dCE}}(f)^2/8 \leq \Delta_{\mathrm{wCE}}(f;\mathcal{L}_1) \leq 2\cdot\Delta_{\mathrm{dCE}}(f).$$

This form of weighted calibration error was previously known as *weak calibration* (Kakade and Foster, 2008) and *smooth calibration* (Gopalan et al., 2022). The connection to 1-Lipschitz weighting functions lends further insight into why dCE may not be the most useful metric for decision

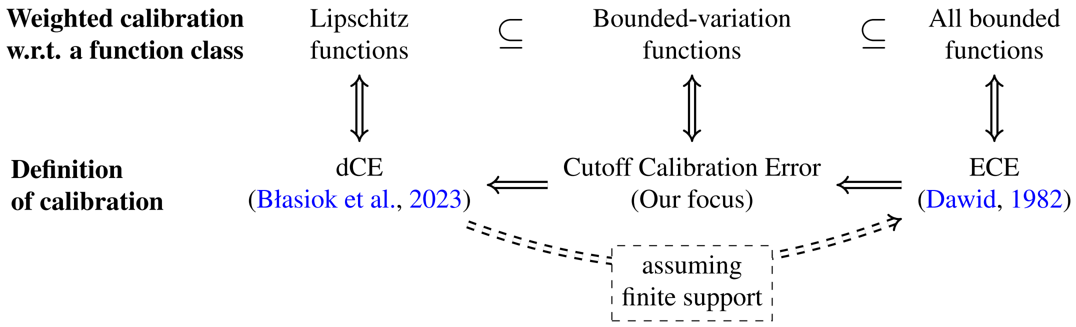
Weighted calibration w.r.t. a function class
Lipschitz functions
⊆
Bounded-variation functions
⊆
All bounded functions
Definition of calibration
dCE (Błasiok et al., 2023)
Cutoff Calibration Error (Our focus)
ECE (Dawid, 1982)
assuming finite support

Figure 2: Relating Cutoff Calibration Error to ECE and dCE. Arrows—of the form $\Delta_1 \Rightarrow \Delta_2$—signify that a calibration metric Δ_2 can be bound in terms of Δ_1.

making: since indicator functions are not Lipschitz, dCE is not compatible with hard thresholding rules. We further explore this issue in Section 3.

On the opposite end of the spectrum, $\Delta_{\text{ECE}}(f)$ is exactly equal to the weighted calibration error using all bounded weighting functions $\Delta_{\text{wCE}}(f; [-1,1]^{[0,1]})$. This observation suggests a challenge of working with ECE: the corresponding function class is too large to be able to estimate this quantity in general. We further explore this issue in Section 4.2.

Finally, note that $\Delta_{\text{Cutoff}}(f)$ is already in the form of a weighted calibration error, where the weighting functions are indicator functions over intervals. Since the objective

$$\mathbb{E}[w(f(X))(Y - f(X))]$$

in weighted calibration error is linear in w, we can replace this class with its convex hull, leading to a much richer class.

Proposition 4 *Let $\mathcal{B}_M$ be the class of all functions on $[0,1] \to [-1,1]$ whose total variation is at most $M \in \mathbb{R}^+$. Then, for $M \geq 2$,*

$$\Delta_{\text{Cutoff}}(f) \leq \Delta_{\text{wCE}}(f; \mathcal{B}_M) \leq (M+3)\Delta_{\text{Cutoff}}(f).$$

This result offers a concrete sense in which Cutoff Calibration Error achieves a useful middle ground between dCE and ECE. The class of bounded variation functions has relatively low complexity, but it is sufficiently expressive to include indicator functions over intervals.

3. Actionable calibration: Implications for binary decision theory

We now turn to analyzing calibration measures through the lens of decision theory, examining what guarantees they provide about forecast quality in practice. In this section, we focus on the binary decision setting, i.e., $Y \in \{0,1\}$. In Appendix B, we demonstrate that analogous results hold for a sign-testing loss function when $Y \in [0,1]$, and even all bounded proper scoring rules when $Y \in \{0,1\}$, under mild regularity assumptions.

3.1. The simple binary decision setting

Consider a simple but fundamental scenario where a decision-maker must choose between two actions (like canceling or proceeding with a flight) based on a probability forecast $f(X)$ of some $Y \in \{0,1\}$. The decision-maker's loss function is determined by their relative tolerance for false positives and false negatives. We write their loss function as $\ell_{\mathrm{bd}} : \{0,1\} \times \{0,1\} \times [0,1] \to [0,1]$, where "bd" refers to "binary decision" loss. For some $\tau \in [0,1]$, the loss is

$$\ell_{\mathrm{bd}}(Y, \widehat{Y}; \tau) = \tau(1-Y)\widehat{Y} + (1-\tau)Y(1-\widehat{Y}),$$

where $\widehat{Y} \in \{0,1\}$ encodes our preferred action if we knew Y.

Given an observed forecast $f(X)$, the Bayes decision rule is given by $\mathbf{1}_{\{\mathbb{E}[Y|f(X)] \geq \tau\}}$. That is to say, the expected loss is minimized when the resources are deployed ($\widehat{Y} = 1$) if and only if the calibrated forecast probability is above τ. In practice, $\mathbb{E}[Y \mid f(X)]$ is unknown, so decision-makers naturally resort to the following plug-in decision rule:

$$\widehat{Y} = \mathbf{1}_{\{f(X) \geq \tau\}}.$$

But, if f is highly miscalibrated, this might be a poor choice of decision rule.

Examining this example, we can see that we are asking whether the decision rule $\mathbf{1}_{\{f(X) \geq \tau\}}$ is (nearly) as good as some other decision rule obtained by transforming f, $\mathbf{1}_{\{h(f(X)) \geq \tau\}}$ (observe that the Bayes rule can be written in this form by taking $h(t) = \mathbb{E}[Y \mid f(X) = t]$). We are particularly interested in whether this favorable property is ensured for functions f that satisfy $\Delta(f) \approx 0$, given some particular choice of calibration measure Δ. This is what we mean by saying that Δ is *actionable*: we are asking whether low miscalibration ($\Delta(f) \approx 0$) implies that the plug-in decision rule is nearly optimal relative to some class of possible modifications of the rule—that is, some set of transformations $\mathcal{H}$.

3.2. Comparing calibration measures through decision guarantees

We now show that the three calibration measures exhibit strikingly different relationships with decision quality. At a high level, Distance from Calibration (dCE) can be arbitrarily small even when a forecast leads to substantially suboptimal decisions. Expected Calibration Error (ECE) provides the strongest guarantee: it bounds the difference between achieved and optimal risk under any transformation of the forecast. Cutoff Calibration Error strikes an appealing middle ground: while not matching the fully general guarantee for ECE, it ensures near-optimal decisions when we restrict attention to the natural class of monotonic decision rules.

Below, we present an example showing that dCE is not guaranteed to be actionable in the binary decision setting, following a similar result from Hu and Wu (2024). Notably, our example is adapted from an example used in Błasiok et al. (2023) to motivate their definition of dCE, since they find it undesirable for a calibration metric to change dramatically under small perturbations in the forecasts. Our example warns that continuity comes at the expense of decision-theoretic guarantees.

Before presenting the example, we first define a measure of risk for the plug-in decision rule:

$$R_{\mathrm{bd}}(f; \tau) := \mathbb{E}[\ell_{\mathrm{bd}}(Y, \mathbf{1}_{\{f(X) \geq \tau\}}; \tau)]. \tag{2}$$

Example 1 *Suppose* $X \sim \mathrm{Bernoulli}(0.75)$ *and* $Y \mid X \equiv X$. *Observe that the constant function* $f(x) := 0.75$ *is perfectly calibrated, so* $\Delta_{\mathrm{dCE}}(f) = \Delta_{\mathrm{ECE}}(f) = 0$. *On the other hand, for a* *slightly perturbed function $\tilde{f}(x) := 0.75 + \epsilon - 2\epsilon x$, we have $\Delta_{\mathrm{dCE}}(\tilde{f}) \leq \epsilon$ and $\Delta_{\mathrm{ECE}}(\tilde{f}) = \frac{3}{8} + \epsilon$. In this case, dCE remains small for small values of ϵ, but ECE is bounded away from zero.*

Suppose $\tau = 0.75$ and we only have access to $\tilde{f}(X)$. Then, the Bayes decision rule will be to use $\widehat{Y} = \mathbf{1}_{\{\mathbb{E}[Y|\tilde{f}(X)] \geq 0.75\}}$. If we use $\mathbf{1}_{\{\tilde{f}(X) \geq 0.75\}}$ in place of the Bayes decision rule, then the risk $R_{\mathrm{bd}}(\tilde{f}(X); \tau)] = \frac{3}{8}$. However, *the Bayes decision rule actually has 0 risk, since in fact $Y = \mathbf{1}_{\{\mathbb{E}[Y|\tilde{f}(X)] \geq 0.75\}}$. Therefore, while $\Delta_{\mathrm{dCE}}(\tilde{f})$ can be arbitrarily small, the discrepancy in losses between the plug-in decision rule and the oracular Bayes decision rule remains bounded away from 0.*

From this example, we can see that a small dCE error, $\Delta_{\mathrm{dCE}}(f) \approx 0$, does not necessarily ensure that the forecast f is actionable.[2]

In contrast, when ECE is small, we can directly guarantee that the risk associated with f is close to the best possible wrapper $h \circ f$. We often refer to this difference as the "*risk gap*."

Proposition 5 *For any $\tau \in [0, 1]$,*

$$R_{\mathrm{bd}}(f; \tau) - \inf_{h:[0,1]\to[0,1]} R_{\mathrm{bd}}(h \circ f; \tau) \leq \Delta_{\mathrm{ECE}}(f).$$

The proof of this follows immediately after the following rearrangement.

Lemma 6 *Let $h : [0, 1] \to [0, 1]$ be arbitrary. Then,*

$$\begin{aligned} &R_{\mathrm{bd}}(f; \tau) - R_{\mathrm{bd}}(h \circ f; \tau) \\ &\qquad = \mathbb{E}[(Y - \tau)\mathbf{1}_{\{f(X)<\tau\}}\mathbf{1}_{\{h(f(X))\geq\tau\}}] + \mathbb{E}[(\tau - Y)\mathbf{1}_{\{f(X)\geq\tau\}}\mathbf{1}_{\{h(f(X))<\tau\}}]. \end{aligned} \tag{3}$$

Proposition 5 also follows from a more general result in Hu and Wu (2024), where instead of considering our specific loss function they worked with all bounded loss functions that are proper scoring rules. They propose a measure of calibration that bounds this risk gap (often termed "swap regret" in game theory). Their proposed definition of calibration ends up resembling the right hand side of (3), with a supremum taken over h and τ, using steps found also in Kleinberg et al. (2023); Li et al. (2022). While their proposed measure of calibration is $\leq \Delta_{\mathrm{ECE}}(f)$, it is also $\geq \Delta_{\mathrm{ECE}}(f)^2$ (Hu and Wu, 2024, Theorem 4.1). Unfortunately, in the setting of a continuous random variable $f(X)$ that is the focus of our work, this means that the proposed calibration measure of Hu and Wu (2024) is not testable, like ECE.

In parallel, Kleinberg et al. (2023) leverage similar observations and propose a calibration metric that is in fact testable. However, its guarantee is for external regret, which compares against a non-personalized treatment policy. Thus, the metric proposed in Kleinberg et al. (2023) is not actionable in the sense that we define here in our work. Relatedly, Qiao and Zhao (2025) consider decision-theoretic guarantees from a different perspective, including benchmarking against non-personalized treatment policies.

[2] Interestingly, there is a certain sense in which dCE is "close" to being actionable: adding a small amount of noise to a forecast $f(X)$ with small dCE will lead to an actionable rule. In fact, though, this is because $f(X) +$ noise is then guaranteed to have small ECE (Blasiok and Nakkiran, 2024). In other words, it is possible to build a low-ECE (and therefore, actionable) forecast by using the low-dCE rule $f(X)$ as a starting point—but $f(X)$ is not itself actionable, and decisions made by thresholding $f(X) +$ noise might be arbitrarily different from decisions made using $f(X)$.

In reality, the best possible wrapper of $f(x)$ (namely, $\mathbb{E}[Y \mid f(X) = f(x)]$) can be very non-smooth, especially when $f(X)$ follows a continuous distribution. The lack of smoothness is not just an abstract concern. For example, no one would implement a decision rule of only bringing your umbrella if the forecast $f(X)$ predicts between a 20% and 50% chance of rain. Yet, an oracle h does not preclude the possibility of suggesting users implement a decision rule that defies common sense. Therefore, the risk gap studied in Proposition 5 is benchmarking against an unrealistic standard.

A natural shape constraint on h that comports with common sense is that it be *monotonically non-decreasing*. By constraining ourselves to monotonic wrappers around f, we can upper bound the *monotone* risk gap by Cutoff Calibration Error.

Proposition 7 *For any* $\tau \in [0, 1]$,

$$R_{\mathrm{bd}}(f;\tau) - \inf_{\substack{h:[0,1]\to[0,1] \\ \text{monotone}}} R_{\mathrm{bd}}(h \circ f;\tau) \leq 2\Delta_{\mathrm{Cutoff}}(f).$$

An alternate way of proving this result, up to constant factors, is to use the results of Okoroafor et al. (2025, Lemmas 3.2 and 3.5), who examine this question from the perspective of omniprediction (Gopalan et al., 2021). The salient fact in both our proof and one leveraging results in Okoroafor et al. (2025) is that pre-images of super-level sets of univariate monotone functions are half-spaces.

While we state Proposition 7 just for binary decision loss, a corollary is that we can bound the risk gaps for arbitrary bounded proper scoring rules when $Y \in \{0, 1\}$ (Appendix B). This corollary follows from the Schervish representation (Schervish, 1989) of proper scoring rules, which demonstrates that proper scoring rules are a mixture of binary decision loss across different τ values.

3.3. Experiment

Set-up To demonstrate how Cutoff Calibration Error satisfies our *actionable* criterion while dCE does not, we consider the following simulation setting.[3] Let $\alpha \in [0, 1]$ be a parameter, and define

$$\mathbb{E}[Y \mid X] = \alpha(1 - 2X)^2 + (1 - \alpha)X.$$

This expression is a convex combination between a parabola (symmetric about 0.5) and the $y = x$ line. For each of our 100 simulation runs, we draw $\alpha \sim \mathrm{Uniform}(0, 1)$ and then sample each $X_i \sim \mathrm{Uniform}(0, 1)$ independently. Each Y_i is drawn from a Bernoulli distribution with mean $\mathbb{E}[Y_i \mid X_i]$. We fit a univariate logistic regression $f : [0, 1] \to [0, 1]$ on $\{(X_i, Y_i)\}_{i=1}^n$, with $n = 500$. Notably, as α increases, the logistic model is increasingly misspecified, as $\mathbb{E}[Y \mid X]$ gains curvature. We evaluate our binary decisions using (2), with $\tau = 0.35$.

Results Figure 3 shows how the risk gap and the monotone risk gap relate to $\Delta_{\mathrm{Cutoff}}(f)$, $\Delta_{\mathrm{ECE}}(f)$, and $\Delta_{\mathrm{dCE}}(f)$. We observe that small values of $\Delta_{\mathrm{ECE}}(f)$ and $\Delta_{\mathrm{Cutoff}}(f)$ correlate with lower risk gaps, guaranteeing that there is no wrapper and monotonic wrapper (respectively) that would meaningfully reduce the expected loss in the binary decision problem. In contrast, $\Delta_{\mathrm{dCE}}(f)$ has a noisy—and at times negligible—relationship with either gap. Thus, even when $\Delta_{\mathrm{dCE}}(f)$ is small, there can exist a wrapper function that would have substantially reduced the expected loss of one's decision. Consequently, the decision-theoretic guarantees with ECE and Cutoff Calibration Error (see Proposition 5 and Proposition 7) are supported by the results of this simulation setting, but dCE does not enjoy these guarantees, neither theoretically nor in the simulation results.

[3] Relevant code is available at https://github.com/rrross/CutoffCalibration

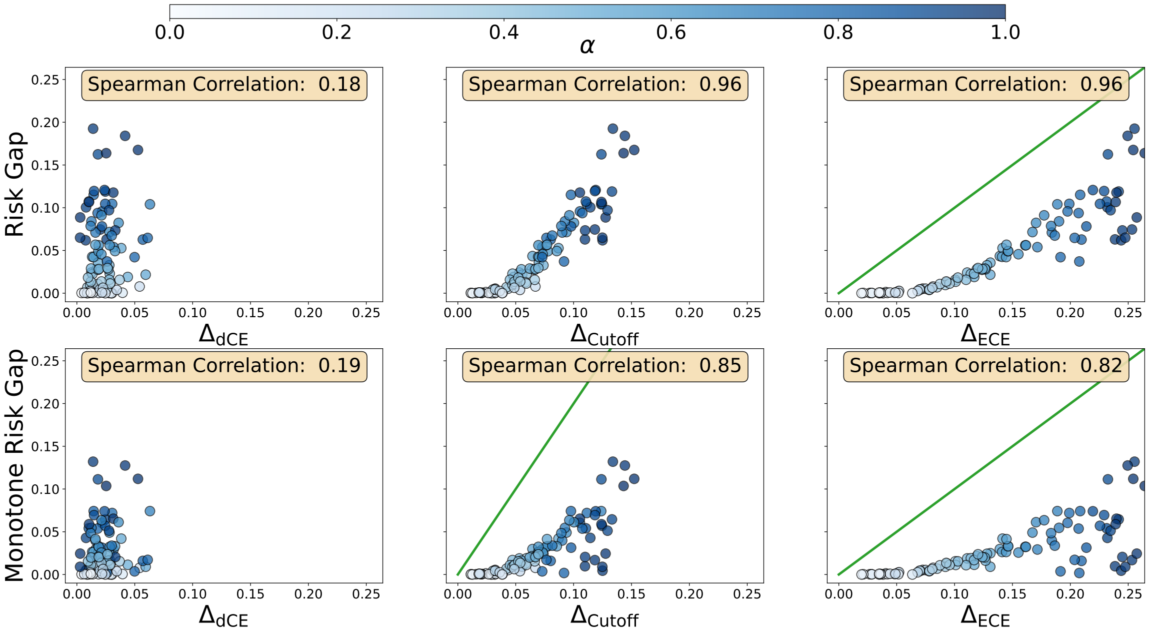

Figure 3: $\Delta_{\text{Cutoff}}(f)$ and $\Delta_{\text{ECE}}(f)$ being small corresponds to smaller risk gaps (both standard and monotone). In contrast, when $\Delta_{\text{dCE}}(f)$ is small, there is no guarantee that the risk gaps will also be small. Green lines denote upper bounds guaranteed by Proposition 5 and Proposition 7. Larger α values correspond to more model-misspecification in our forecast model f. Calibration measures $\Delta_{\text{dCE}}(f), \Delta_{\text{Cutoff}}(f), \Delta_{\text{ECE}}(f)$ are oracle quantities in the sense that we calculate them using knowledge of $\mathbb{E}[Y \mid X]$; see Appendix B.

4. Testable calibration: Distribution-free guarantees

In order for a calibration metric to be useful for decision-making in practice, as discussed in Section 3, we need to choose a metric Δ that is *actionable*. But this is not sufficient: we also need to be able to verify empirically that $\Delta^P(f_n) \approx 0$,[4] in order to know whether our particular trained model f_n can then be safely used for decision-making. In other words, for a small constant $c > 0$,

> Given a sample of data points drawn i.i.d. from P, can we fit a function f_n to the data, in such a way that guarantees $\Delta^P(f_n) \leq c$ with high probability?

Of course, this is always possible with a trivial solution: we can simply return a constant function $f_{n,\text{const}}(x) \equiv \hat{\mu}_Y$ where $\hat{\mu}_Y$ is the sample mean of Y (since we can always estimate the mean $\mathbb{E}[Y]$ with error $\mathcal{O}_P(n^{-1/2})$, this will ensure that $\Delta^P(f_{n,\text{const}})$ is low). However, we would ideally want to use a state-of-the-art fitted model f_n, not some overly simple model such as a constant function. We can therefore ask a related question:

[4] When we write $\Delta^P(f_n)$, we proceed as if f_n were fixed. We use the P superscript to emphasize that the expectation is only taken over P, not over the data sampled from P^n and used to train f_n (that is, the expectation is computed conditional on f_n). In particular for a fixed function f, $\Delta(f) = \Delta^P(f)$, since conditioning on the fixed function f would have no effect.

> (Testability.) Given a fixed function f and a sample of data points drawn i.i.d. from P, can we test the hypothesis $\Delta(f) \leq c$?

To see how this relates to the previous question, we can consider the following workflow, given a data set of size n:

- Using $n/2$ data points, train an initial predictive model $f_{n,\text{init}}$.
- Using the remaining $n/2$ data points, test whether $\Delta^P(f_{n,\text{init}}) \leq c$. If yes, then return $f_n = f_{n,\text{init}}$. If not, then return a constant function, $f_n = f_{n,\text{const}}$, for $f_{n,\text{const}}(x) \equiv \hat{\mu}_Y$.

In other words, if Δ is testable, then we are able to return a model f_n that is (with high probability) guaranteed to satisfy $\Delta^P(f_n) \leq c$.

In this section, at a high level, our results show that dCE and Cutoff Calibration Error are both testable—and therefore, it is possible to use data to produce a fitted model f_n that is guaranteed to satisfy these calibration metrics. On the other hand, ECE is testable only for piecewise constant functions f and is not testable more generally; equivalently, any procedure that returns a f_n that is guaranteed to have low ECE, must therefore return a f_n that is (usually) piecewise constant, which immediately excludes many standard models such as logistic regression (or Platt scaling).

4.1. Testability of dCE and Cutoff Calibration Error

We will next see that Cutoff Calibration Error and dCE are both testable: it is possible to test the hypothesis $\Delta_{\text{Cutoff}}(f) \leq c$ or $\Delta_{\text{dCE}}(f) \leq c$. In particular, as we will formalize below, we can construct practical procedures to return a function f_n guaranteed to satisfy a bound on calibration error.

We begin by considering Cutoff Calibration Error. First, we establish that the natural estimator for $\Delta_{\text{Cutoff}}(f)$ is consistent.

Proposition 8 *Suppose $(X_i, Y_i) \overset{iid}{\sim} P$ for $i = 1, \ldots, n$ and $Y \in [0, 1]$. Let $\widehat{\Delta}_{\text{Cutoff}}(f) : \mathcal{F} \to [0, 1]$ be a plug-in estimator for Cutoff Calibration Error defined as*

$$\widehat{\Delta}_{\text{Cutoff}}(f) = \sup_{I:\text{ interval}} \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i)) \mathbf{1}_{\{f(X_i) \in I\}} \right|.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\left| \widehat{\Delta}_{\text{Cutoff}}(f) - \Delta_{\text{Cutoff}}(f) \right| \leq \frac{20 + \sqrt{2 \log(1/\delta)}}{\sqrt{n}}.$$

In particular, we can therefore use this estimator to test whether cutoff calibration error is low. In other words, given some threshold $c > 0$, if $\widehat{\Delta}_{\text{Cutoff}}(f) \leq c - \frac{20+\sqrt{2\log(1/\delta)}}{\sqrt{n}}$ then we can certify (with $1 - \delta$ confidence) that $\Delta_{\text{Cutoff}}(f) \leq c$—and in particular, this procedure may have nontrivial power as soon as $n \gtrsim \frac{\log(1/\delta)}{c^2}$. An analogous result on testability for a closely related calibration measure is provided by Okoroafor et al. (2025, Lemma C.2).

More generally, however, we may be interested in providing a procedure that is *guaranteed* to offer Cutoff Calibration Error bounded by some threshold c, rather than *testing* whether Cutoff

Calibration Error is bounded by c (or not). We can also use the result above to provide a procedure that offers such a guarantee. Given a data set $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P$, a threshold $c > 0$ and $\delta \in (0,1)$, and a model fitting algorithm $\mathcal{A}$ (e.g., a neural net),

- Train a model $f_{n,\text{init}}$ using the first half of the data: $f_{n,\text{init}} = \mathcal{A}\big((X_1, Y_1), \ldots, (X_{\lceil n/2 \rceil}, Y_{\lceil n/2 \rceil})\big)$.
- Estimate the calibration error of $f_{n,\text{init}}$ using the second half of the data:

$$\widehat{\Delta}_{\text{Cutoff}}(f_{n,\text{init}}) = \sup_{I:\text{ interval}} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=\lceil n/2 \rceil + 1}^{n} (Y_i - f(X_i)) \mathbf{1}_{\{f_{n,\text{init}}(X_i) \in I\}} \right| .$$

- If $\widehat{\Delta}_{\text{Cutoff}}(f_{n,\text{init}}) \leq c - \frac{20 + \sqrt{2\log(1/\delta)}}{\sqrt{\lfloor n/2 \rfloor}}$, return $f_n = f_{n,\text{init}}$, else return $f_n = f_{n,\text{const}}$ where

$$f_{n,\text{const}}(x) \equiv \frac{1}{\lfloor n/2 \rfloor} \sum_{i=\lceil n/2 \rceil + 1}^{n} Y_i .$$

We then have the following guarantee:

Corollary 9 *For any distribution P, and any model fitting algorithm $\mathcal{A}$, the procedure described above satisfies*

$$\mathbb{P}\{\Delta^P_{\text{Cutoff}}(f_n) \leq c\} \geq 1 - 2\delta$$

as long as $c \geq \sqrt{\frac{\log(1/\delta)}{2\lfloor n/2 \rfloor}}$.

In particular, note that the result holds uniformly over any distribution P, and the procedure does not require any knowledge of P—that is, this is a distribution-free guarantee. Moreover, there are no restrictions on the model fitting algorithm $\mathcal{A}$, and so the returned model f_n may be quite complex and adaptive to the data, if we choose our algorithm well for the task at hand.

Next, consider dCE. In fact, since $\Delta_{\text{dCE}}(f) \leq 3\sqrt{\Delta_{\text{Cutoff}}(f)}$ by Proposition 2, this means that any guarantee of an upper bound on $\Delta_{\text{Cutoff}}(f)$ (as in Corollary 9 above, say) immediately yields a bound on $\Delta_{\text{dCE}}(f)$, as well. Relatedly, Błasiok et al. (2023, Section 9) establish that a calibration measure related to dCE can be estimated consistently using an i.i.d. sample, with error $\mathcal{O}_P(n^{-1/2})$, which leads to lower and upper bounds on dCE, as well.

4.2. Hardness result: ECE is not testable

In contrast to dCE and Cutoff Calibration Error, the ECE calibration metric is not testable in the distribution-free setting; there is no estimator of $\Delta_{\text{ECE}}(f)$ that is guaranteed to be consistent for any f and uniformly over any distribution P. In particular, this means that we cannot construct a two-stage procedure in the style of the procedure described in Section 4.1 for Cutoff Calibration Error and dCE. More strongly, we will now see that *any* mechanism for returning a function f_n guaranteed to have low ECE is necessarily forced to return an output that is usually piecewise constant; we cannot output, say, a continuous and strictly increasing f_n (as would be the case for procedures such as post-hoc calibration via Platt scaling), if we require a distribution-free guarantee on ECE.

To state the result, we first need some additional notation. First, for any $\gamma \in [0,1]$ and any distribution P on $\mathcal{X} \times [0,1]$ we define

$$\sigma^2_\gamma(P) = \inf\left\{\mathbb{E}_P\left[\mathrm{Var}(Y \mid X) \cdot \mathbf{1}_{\{X \in A\}}\right] \ : \ A \subseteq \mathcal{X}, \mathbb{P}_P(X \in A) \geq \gamma\right\}.$$

We can think of this quantity as an "effective variability" of P—in particular if $\mathrm{Var}(Y \mid X) \geq \sigma^2 > 0$ almost surely, then $\sigma^2_\gamma(P) \geq \sigma^2$ for any γ. Next, we also define a notion of "effective support size":

$$S_\gamma(f, P) = \inf\left\{k \geq 1 : \mathbb{P}_P\big(f(X) \in \{t_1, \dots, t_k\}\big) \geq 1 - \gamma \text{ for some } t_1, \dots, t_k \in [0,1]\right\},$$

or $S_\gamma(f,P) = +\infty$ if the set is empty. If $f(X)$ takes at most k many values (under $X \sim P$), then $S_\gamma(f,P) \leq k$ for any γ.[5]

Theorem 10 *Fix any $c, \delta > 0$ and any sample size $n \geq 1$. Let*

$$\mathcal{A} : (\mathcal{X} \times [0,1])^n \to \{\textit{ measurable functions } f : \mathcal{X} \to [0,1]\}$$

be any procedure that inputs a data set $(X_1, Y_1), \dots, (X_n, Y_n)$, and returns a fitted function $f_n = \mathcal{A}\big((X_1,Y_1),\dots,(X_n,Y_n)\big)$.

If $\mathcal{A}$ satisfies the distribution-free guarantee

$$\mathbb{P}_{P^n}\{\Delta^P_{\mathrm{ECE}}(f_n) \leq c\} \geq 1 - \delta \text{ for all distributions } P,$$

then it holds that

$$\mathbb{P}_{P^n}\left\{S_\gamma(f_n, P) \leq n^2\right\} \geq 1 - \frac{2e(c + \delta + n^{-1})}{\sigma^2_\gamma(P)} \text{ for all distributions } P.$$

If c and δ are small (i.e., $\mathcal{A}$ offers a meaningful guarantee on ECE), then for any P with an effective variance bounded away from zero, we must have $S_\gamma(f_n, P) \leq n^2$ with high probability—that is, f_n must be a function that is piecewise constant on most of the domain, with effective support at most $\mathcal{O}(n^2)$.

We emphasize that this result applies to *any* procedure $\mathcal{A}$—for instance, $\mathcal{A}$ can be the outcome of training a model on part of the data, and then using a post-hoc calibration procedure such as Platt scaling on an additional batch of data. In other words, this result implies that post-hoc calibration does not have any meaningful distribution-free guarantee in terms of the ECE, unless we use a procedure that returns a piecewise constant output. Thus, ECE does not satisfy our testability criterion for a very broad class of models.

Related work by Gupta et al. (2020) examines this type of question from an asymptotic perspective, finding that any post-hoc calibration procedure satisfying an asymptotic guarantee on ECE cannot return injective functions. In Appendix C, we will compare our result to this existing work and see how Gupta et al. (2020)'s asymptotic hardness result is implied by Theorem 10.

[5] As for calibration error, when the function f_n is data-dependent, we treat f_n as fixed for the purpose of calculating $S_\gamma(f_n, P)$—that is, $\mathbb{P}_P(f_n(X) \in \{t_1, \dots, t_k\})$ should be interpreted as a probability calculated with respect to an independent draw of X, when we condition on the trained model f_n.

Estimating ECE via binning? In practice, it is common to estimate ECE with a binning-based approach: that is, given a function f and a partition $[0,1] = A_1 \cup \cdots \cup A_N$ (with N small relative to sample size n), we use the available data estimate the error

$$|\mathbb{E}[Y \mid f(X) \in A_i] - \mathbb{E}[f(X) \mid f(X) \in A_i]|,$$

and aggregate over bins $i = 1, \ldots, N$, to estimate $\Delta_{\mathrm{ECE}}(f)$ (Błasiok et al., 2023). In fact, this type of binned approximation (known as the binned ECE) is actually estimating the ECE of *the binned approximation* to the function f (that is, a function g that is piecewise constant, taking the value $g(x) = \mathbb{E}[f(X) \mid f(X) \in A_i]$ for each $x \in A_i$)—and without further assumptions, it is possible to have $\Delta_{\mathrm{ECE}}(g) \approx 0$ even when $\Delta_{\mathrm{ECE}}(f)$ is large. This challenge is closely connected to our hardness result, Theorem 10.

5. Methodological insights on post-hoc calibration

In this section, we examine the implications of Cutoff Calibration Error on post-hoc calibration. Post-hoc calibration, in short, aims to complement a pre-fit forecast model f by using data to find an $h_n : [0,1] \to [0,1]$ such that $h_n \circ f$ achieves small calibration error.

5.1. Isotonic regression enjoys uniform asymptotic control of Cutoff Calibration Error

A popular calibration method is to use isotonic regression on the observed outcomes Y as a monotone function of the forecasts $f(X)$ (Zadrozny and Elkan, 2002). The isotonic constraint is natural for reasons similar to those outlined in Section 3. Namely, one may be concerned that a non-monotonic h_n would be the result of over-fitting. Moreover, even if there were true non-monotonic regions in $\mathbb{E}[Y \mid f(X)]$ as a function of $f(X)$, one may view correcting those errors as more in the purview of creating a better base forecast f.

Proposition 11 *Suppose $(X_i, Y_i) \overset{iid}{\sim} P$ for $i = 1, \ldots, n$. Assume $Y_i \in [0,1]$ and f is fixed. Let h_n refer to running isotonic regression on $\{(f(X_i), Y_i)\}_{i=1}^n$. Fix $\delta > 0$. Then,*

$$\Delta^P_{\text{Cutoff}}(h_n \circ f) \leq \frac{30 + 2\sqrt{2\log(2/\delta)}}{\sqrt{n}}$$

with probability at least $1 - \delta$.

5.2. Platt scaling does not necessarily calibrate, even in dCE

Platt scaling (Platt et al., 1999) is a popular post-hoc calibration method which essentially consists of running a univariate logistic regression of observed Y values on previous forecasts $f(X)$.[6] The issue, in short, is that there is no guarantee that Platt scaling will converge asymptotically to a calibrated function. Since $\Delta_{\mathrm{dCE}}(f) = 0$ if and only f is perfectly calibrated, this implies that Platt scaling cannot provide strong asymptotic calibration on dCE, ECE, and Cutoff Calibration Error. In Figure 4, we exhibit an example of a data-generation process for which the population Platt scaling

[6] For finite samples, Platt scaling minimizes logistic loss with $\tilde{Y}_i = Y_i \cdot \frac{\sum_{i=1}^n Y_i+1}{\sum_{i=1}^n Y_i+2} + (1-Y_i)\frac{1}{\sum_{i=1}^n (1-Y_i)+2}$ in place of Y_i to improve empirical performance.

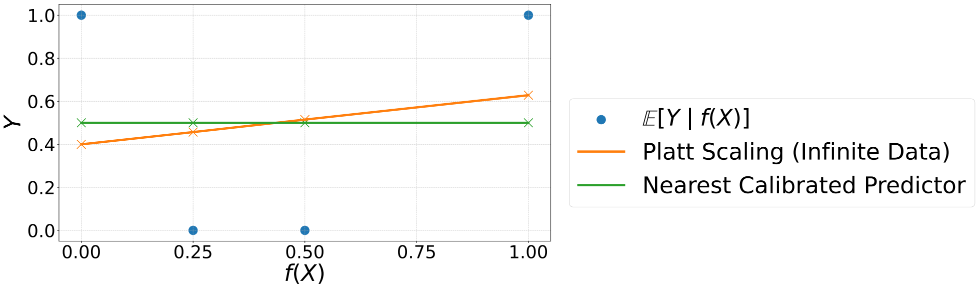

Figure 4: **Platt scaling does not guarantee small dCE even with infinite data.** In our example, we assume $f(X) \sim \mathrm{Uniform}\{0, 0.25, 0.5, 1\}$.

algorithm yields an $h : [0, 1] \to [0, 1]$ (attained by optimizing logistic loss in expectation instead of over samples) that has non-zero dCE: $\Delta_{\mathrm{dCE}}(h \circ f) > 0$.

In Appendix D, we propose a modified Platt scaling algorithm that asymptotically controls Cutoff Calibration Error without data splitting, achieving the same theoretical guarantee as isotonic regression for Cutoff Calibration Error (Proposition 11).

6. Discussion

There are several promising research directions and intriguing connections that stem from our findings. First, while our work focuses on the i.i.d. setting, we think our findings may hold ramifications for the branch of calibration research that focuses on sequential prediction and could potentially inform the design of new online calibration algorithms. Second, there may be connections between our work and Sahoo et al. (2021), who also focus on "threshold" decision rules, although in a different setting than us, where forecasters provide a CDF estimate for a real-valued output. Third, the monotonic constraint used in our decision-theoretic guarantees may provide intriguing results in other contexts, such as that of Roth and Shi (2024), who examine the impact of the dimensionality of a prediction space on swap regret and calibration. We also note that our focus on interval regions of the prediction space is related to the 1-dimensional special case in Roth and Shi (2024), which builds on the work in Noarov et al. (2023) in showing that actionability is related to performance within convex subsets of the prediction space. Finally, while the decision-theoretic guarantees of Cutoff Calibration Error are broadest when $Y \in \{0, 1\}$, we believe that there may be a more general guarantee related to Cutoff Calibration Error when $Y \in [0, 1]$.

Acknowledgments

The authors gratefully acknowledge the National Science Foundation via grant DMS-2023109. J.S. and R.F.B. were partially supported by the Office of Naval Research via grant N00014-24-1-2544. J.S. was also partially supported by the Margot and Tom Pritzker Foundation. Z.R. was supported by NSF grant DMS-2413135. R.M.W. was partially supported by the NSF-Simons National Institute for Theory and Mathematics in Biology (NITMB) via grants NSF DMS-2235451 and Simons Foundation MP-TMPS-00005320.

References

Imanol Arrieta-Ibarra, Paman Gujral, Jonathan Tannen, Mark Tygert, and Cherie Xu. Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*, 23(351):1–54, 2022.

Jaroslaw Blasiok and Preetum Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XwiA1nDahv.

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

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

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint arXiv:2109.05389*, 2021.

Parikshit Gopalan, Michael P Kim, Mihir A Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, pages 3193–3234. PMLR, 2022.

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

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.

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

Lunjia Hu and Yifan Wu. Predict to minimize swap regret for all payoff-bounded tasks. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 244–263, 2024. doi: 10.1109/FOCS61266.2024.00024.

Sham M Kakade and Dean P Foster. Deterministic calibration and Nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

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

Andreĭ Nikolaevich Kolmogorov and Sergeí Vasil'evich Fomin. *Introductory real analysis*. Courier Corporation, 1975.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

Gregory F Lawler. *Random walk and the heat equation*, volume 55. American Mathematical Soc., 2010.

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

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. *arXiv preprint arXiv:2310.17651*, 2023.

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

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

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

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

Aaron Roth and Mirah Shi. Forecasting for swap regret for all downstream agents. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 466–488, 2024.

Roshni Sahoo, Shengjia Zhao, Alyssa Chen, and Stefano Ermon. Reliable decisions with threshold calibration. *Advances in Neural Information Processing Systems*, 34:1831–1844, 2021.

Mark J Schervish. A general method for comparing probability assessors. *The annals of statistics*, 17(4):1856–1879, 1989.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

Appendix A. Relationships among calibration measures: proofs and examples

A.1. dCE and ECE are equivalent in the discrete setting

In this section, we show that in the case where the prediction $f(X)$ is a discrete random variable, dCE and ECE are essentially equivalent.

Proposition 12 *Defining* $b_f = \operatorname{ess\,inf}\{|f(x) - f(x')| : x, x' \in \mathcal{X}, f(x) \neq f(x')\}$,

$$\Delta_{\mathrm{dCE}}(f) \geq \frac{b_f}{b_f + 1} \cdot \Delta_{\mathrm{ECE}}(f).$$

Proof Suppose $f(X)$ has finite support. We can write $f(X) \in \{u_1, \dots, u_m\}$ for some distinct values $u_1, \dots, u_m$, with $\min_{i \neq j} |u_i - u_j| \geq b_f$ by assumption.

Fix any $\epsilon > 0$. By definition of $\Delta_{\mathrm{dCE}}(f)$, there exists some perfectly calibrated function g such that

$$\mathbb{E}[|f(X) - g(X)|] \leq \Delta_{\mathrm{dCE}}(f) + \epsilon.$$

Define also a probability vector $p(g(X))$ with entries

$$p_i(g(X)) = \mathbb{P}\{f(X) = u_i \mid g(X)\},$$

for $i \in [m]$. Moreover, define a random index $i^*(g(X)) \in \arg\max_{i \in [m]} p_i(g(X))$. That is, given $g(X)$, the index $i^*(g(X))$ identifies the (not necessarily unique) most likely value of $f(X)$ in the finite set $\{u_1, \dots, u_m\}$. Define also another random index, $i_{\mathrm{Med}}(g(X))$, satisfying that $u_{i_{\mathrm{Med}}(g(X))}$ is the (not necessarily unique) median of the distribution of $f(X)$ conditional on $g(X)$.

Next, let w be any function taking values in $[-1, 1]$. We need to bound $\mathbb{E}[w(f(X)) \cdot (Y - f(X))]$. First, define

$$\tilde{w}(g(X)) = w(u_{i^*(g(X))}).$$

Next, we calculate

$$\begin{aligned}
\mathbb{P}\{w(f(X)) \neq \tilde{w}(g(X)) \mid g(X)\} &\leq \mathbb{P}\{f(X) \neq u_{i^*(g(X))} \mid g(X)\} \\
&= 1 - p_{i^*(g(X))}(g(X)) \\
&= 1 - \max_{i \in [m]} p_i(g(X)).
\end{aligned}$$

And,

$$\begin{aligned}
\mathbb{E}[|f(X) - g(X)| \mid g(X)] &\geq \inf_{t \in [0,1]} \mathbb{E}[|f(X) - t| \mid g(X)] \\
&= \mathbb{E}[|f(X) - u_{i_{\mathrm{Med}}(g(X))}| \mid g(X)] \\
&\geq \mathbb{E}[b_f \cdot \mathbf{1}_{\{f(X) \neq u_{i_{\mathrm{Med}}(g(X))}\}} \mid g(X)] \\
&= b_f \cdot \left(1 - p_{u_{i_{\mathrm{Med}}(g(X))}}(g(X)\right) \\
&\geq b_f \cdot \left(1 - \max_{i \in [m]} p_i(X)\right),
\end{aligned}$$

where the second step holds since the expected absolute loss is minimized by the median, and the third step holds by our assumption that $\min_{i \neq j} |u_i - u_j| \geq b_f$. Combining these calculations, we have shown that

$$\mathbb{P}\{w(f(X)) \neq \tilde{w}(g(X)) \mid g(X)\} \leq \frac{1}{b_f} \cdot \mathbb{E}[|f(X) - g(X)| \mid g(X)].$$

After marginalizing over $g(X)$, then,

$$\mathbb{P}\{w(f(X)) \neq \tilde{w}(g(X))\} \leq \frac{1}{b_f} \cdot \mathbb{E}[|f(X) - g(X)|] \leq \frac{1}{b_f} \left(\Delta_{\text{dCE}}(f) + \epsilon\right).$$

Finally, we bound ECE. We have

$$\begin{aligned}\mathbb{E}[w(f(X)) \cdot (Y - f(X))] = \mathbb{E}[(w(f(X)) - \tilde{w}(g(X))) \cdot (Y - f(X))] + \\ \mathbb{E}[\tilde{w}(g(X)) \cdot (g(X) - f(X))] + \mathbb{E}[\tilde{w}(g(X)) \cdot (Y - g(X))].\end{aligned}$$

For the first term, since $|(w(f(X)) - \tilde{w}(g(X))) \cdot (Y - f(X))| \leq 1$ holds almost surely, we have

$$\mathbb{E}[(w(f(X)) - \tilde{w}(g(X))) \cdot (Y - f(X))] \leq \mathbb{P}\{w(f(X)) \neq \tilde{w}(g(X))\} \leq \frac{1}{b_f} \left(\Delta_{\text{dCE}}(f) + \epsilon\right).$$

For the second term, since $|\tilde{w}(g(X))| \leq 1$ almost surely,

$$\mathbb{E}[\tilde{w}(g(X)) \cdot (g(X) - f(X))] \leq \mathbb{E}[|f(X) - g(X)|] \leq \Delta_{\text{dCE}}(f) + \epsilon.$$

For the third term, since g is perfectly calibrated,

$$\mathbb{E}[\tilde{w}(g(X)) \cdot (Y - g(X))] = 0.$$

Combining everything, then, we have shown that

$$\mathbb{E}[w(f(X)) \cdot (Y - f(X))] \leq \left(1 + \frac{1}{b_f}\right) \cdot \left(\Delta_{\text{dCE}}(f) + \epsilon\right).$$

Since this holds for all w, and since $\epsilon > 0$ can be taken to be arbitrarily small, this completes the proof. ■

A.2. Proof of Proposition 2: Relationship between dCE, Cutoff Calibration Error and ECE

We break down the series of inequalities in Proposition 2 into the following lemmas.

Lemma 13 *For any function f,*

$$\Delta_{\text{Cutoff}}(f) \leq \Delta_{\text{ECE}}(f).$$

Proof This follows immediately from Section 2.1, which demonstrates that Cutoff Calibration Error and ECE are both instances of wCE and that the complexity class of weight functions for ECE is a superset of the complexity class for Cutoff Calibration Error. ■

Lemma 14 *For any function f,*

$$\Delta_{\text{dCE}}(f)^2/9 \le \Delta_{\text{Cutoff}}(f).$$

Proof Fix $N \in \mathbb{N}$. Partition $[0,1]$ into N equal-length intervals:

$$A_1 = [0, 1/N]; A_2 = (1/N, 2/N]; \ldots; A_N = ((N-1)/N, 1].$$

Define $\psi : [0,1] \to [N]$ as

$$\psi(z) := \sum_{j=1}^{N} j \mathbf{1}_{\{z \in A_j\}},$$

i.e., $\psi(z)$ provides the index of which interval of $A_1, \ldots, A_N$ that z lies in.

Define $g : \mathcal{X} \to [0,1]$ as follows:

$$g(x) := \mathbb{E}[Y \mid \psi(f(X)) = \psi(f(x))].$$

Observe that g is a calibrated:

$$\mathbb{E}[Y \mid g(X)] = \mathbb{E}[\mathbb{E}[Y \mid \psi(f(X))] \mid g(X)] = \mathbb{E}[g(X) \mid g(X)] = g(X).$$

Define a random variable V as

$$V := \mathbb{E}[f(X) \mid \psi(f(X))].$$

We know that $\mathbb{E}|f(X) - V| \le 1/N$ almost surely, since $\forall j \in [N]$ we have

$$\psi(f(X)) = j \implies f(X), V \in A_j \implies |f(X) - V| \le \text{diam}(A_j) = 1/N.$$

Therefore,

$$\mathbb{E}|g(X) - f(X)| \le \mathbb{E}|g(X) - V| + \mathbb{E}|V - f(X)| \le \mathbb{E}|g(X) - V| + 1/N.$$

Focusing on the first term,

$$\begin{aligned}
\mathbb{E}|g(X) - V| &= \mathbb{E}|\mathbb{E}[Y - f(X) \mid \psi(f(X))]| \\
&= \sum_{j=1}^{N} \mathbb{P}\{f(X) \in A_j\} |\mathbb{E}[(Y - f(X)) \mid f(X) \in A_j]| \\
&= \sum_{j=1}^{N} |\mathbb{E}[(Y - f(X)) \mathbf{1}_{\{f(X) \in A_j\}}]| \\
&\le N \Delta_{\text{Cutoff}}(f).
\end{aligned}$$

Therefore, since g is calibrated,

$$\Delta_{\text{dCE}}(f) \le N \Delta_{\text{Cutoff}}(f) + 1/N.$$

Setting $N = \left\lceil \frac{1}{\sqrt{\Delta_{\text{Cutoff}}(f)}} \right\rceil$, we get

$$\Delta_{\text{dCE}}(f) \le 2\sqrt{\Delta_{\text{Cutoff}}(f)} + \Delta_{\text{Cutoff}}(f) \le 3\sqrt{\Delta_{\text{Cutoff}}(f)}.$$

■

A.3. Proof of Proposition 4: Cutoff Calibration and bounded-variation functions

Proof [Proposition 4] The first inequality follows immediately from observing that indicator functions over intervals have a total variation of at most 2 and that, if $g \in \mathcal{B}_M$, then $-g \in \mathcal{B}_M$.

For the second inequality, by Lemma 15, $\exists g_1, g_2 : [0,1] \to [-\frac{1}{2}, \frac{M+1}{2}]$ non-decreasing such that $g = g_1 - g_2$.

Define $\mathcal{G} := \{g : [0,1] \to [-\frac{1}{2}, \frac{M+1}{2}] \mid g \text{ non-decreasing}\}$.

Define $\mathcal{G}' = \{g : [0,1] \to [0,1] \mid g \text{ non-decreasing}\}$.

Therefore,

$$\begin{aligned}
&\sup_{g \in \mathcal{B}_M} \mathbb{E}[(Y - f(X))g(f(X))] \\
&\leq \sup_{g_1, g_2 \in \mathcal{G}} (\mathbb{E}[(Y - f(X))g_1(f(X))] - \mathbb{E}[(Y - f(X))g_2(f(X))]) \\
&= \sup_{g \in \mathcal{G}} \mathbb{E}[(Y - f(X))g(f(X))] + \sup_{g \in \mathcal{G}} \mathbb{E}[(Y - f(X))(-g(f(X)))] \\
&\leq 2 \sup_{g \in \mathcal{G} \cup (-\mathcal{G})} \mathbb{E}[(Y - f(X))g(f(X))] \\
&= 2 \sup_{g \in \mathcal{G}' \cup (-\mathcal{G}')} \mathbb{E}\left[(Y - f(X))\left((M/2 + 1) \cdot g(f(X)) - 0.5 + \mathbf{1}_{\{g \in (-\mathcal{G}')\}}\right)\right] \\
&\leq (M + 2) \sup_{g \in \mathcal{G}' \cup (-\mathcal{G}')} \mathbb{E}[(Y - f(X))g(f(X))] + |\mathbb{E}[Y - f(X)]| \\
&\leq (M + 3) \sup_{g \in \mathcal{G}' \cup (-\mathcal{G}')} \mathbb{E}[(Y - f(X))g(f(X))] \\
&\leq (M + 3)\Delta_{\text{Cutoff}}(f).
\end{aligned}$$

The final inequality follows from observing both that

$$\mathcal{G}' = \overline{\text{conv}}\{g(x) = \mathbf{1}_{\{x \in [t,1]\}} \mid t \in [0,1]\}$$

and that we are taking the supremum of an objective function that is linear with respect to its argument g. More precisely,

$$\begin{aligned}
&\sup_{g \in \mathcal{G}' \cup (-\mathcal{G}')} \mathbb{E}[(Y - f(X))g(f(X))] \\
&= \max\left\{\sup_{g \in \mathcal{G}'} \mathbb{E}[(Y - f(X))g(f(X))], \sup_{g \in \mathcal{G}'} \mathbb{E}[(Y - f(X))(-g(f(X)))]\right\} \\
&= \max\left\{\sup_t \mathbb{E}[(Y - f(X))\mathbf{1}_{\{f(X) \in [t,1]\}}], \sup_t \mathbb{E}[(Y - f(X))(-\mathbf{1}_{\{f(X) \in [t,1]\}})]\right\} \\
&\leq \sup_t |\mathbb{E}[(Y - f(X))\mathbf{1}_{\{f(X) \in [t,1]\}}]|.
\end{aligned}$$

■

Lemma 15 *Suppose $f : [0,1] \to [-1,1]$ and has total variation at most M. Then, $\exists f_1, f_2 : [0,1] \to [-\frac{1}{2}, \frac{M+1}{2}]$ non-decreasing such that $f = f_1 - f_2$.*

Proof We proceed similarly to Kolmogorov and Fomin (1975, Section 32 Theorem 4).

Let $v(x)$ be the total variation of f on $[0, x]$. Observe that $\mathrm{im}(v) \subseteq [0, M]$.

Define $f_1 = 0.5(v + f)$ and $f_2 = 0.5(v - f)$.

We know f_2 is non-decreasing from Kolmogorov and Fomin (1975, Section 32 Theorem 4).

To show f_1 is non-decreasing, let $x \leq y$, where $x, y \in [0, 1]$.

Then,

$$(v + f)(y) - (v + f)(x) = v(y) - v(x) + f(y) - f(x) \geq 0,$$

since $|f(y) - f(x)| \leq v(y) - v(x)$ (Kolmogorov and Fomin, 1975, Section 32 Theorem 4).

Conclude by noting $f = f_1 - f_2$. ■

A.4. Examples that demonstrate tightness

Example where dCE, Cutoff Calibration Error, and ECE are equal Suppose $Y \equiv 0$ and $f(x) \equiv 1$. Then, the only calibrated $g : \mathcal{X} \to [0, 1]$ is $g(x) \equiv 0$. Therefore, $\Delta_{\text{dCE}}(f) = 1$. Similarly, $|\mathbb{E}[(Y - f(X))\mathbf{1}_{\{f(X)\in I\}}]|$ for some interval I is upper bounded by $|\mathbb{E}[(Y - f(X))]| = 1$; therefore, $\Delta_{\text{Cutoff}}(f) = 1$. Finally, $\Delta_{\text{ECE}}(f) = \mathbb{E}|1 - 0| = 1$. Therefore,

$$\Delta_{\text{dCE}}(f) = \Delta_{\text{Cutoff}}(f) = \Delta_{\text{ECE}}(f).$$

Example showing Proposition 12 is tight Let $\mathcal{X} = [0, 1]$, with $X \sim \text{Unif}[0, 1]$ and $Y = \mathbf{1}_{\{X>0.5\}}$. Let $f(x) = 0.5(1 + b) - b \cdot \mathbf{1}_{\{X>0.5\}}$, which satisfies the assumption (i.e., the separation is $\geq b$). Let $g(x) \equiv 0.5$, which is perfectly calibrated. Clearly $\Delta_{\text{dCE}}(f)$ is attained by comparing to g, i.e., $\Delta_{\text{dCE}}(f) = \mathbb{E}|f(X) - g(X)| = 0.5b$. And, $\Delta_{\text{ECE}}(f) = 0.5(1 + b)$, so,

$$\Delta_{\text{ECE}}(f) = \left(1 + \frac{1}{b}\right) \Delta_{\text{dCE}}(f).$$

Example showing the quadratic relationship between dCE and Cutoff Calibration Error Suppose $f(X) \sim \text{Uniform}(0, 1)$. Suppose $\mathbb{E}[Y \mid f(X)] = \frac{i-0.5}{N}$ for $f(X) \in (\frac{i-1}{N}, \frac{i}{N}]$. Then, $\Delta_{\text{Cutoff}}(f) = |\mathbb{E}[(Y - f(X))\mathbf{1}_{\{f(X)\in(0,\frac{0.5}{N}\}}]| = \frac{1}{8N^2}$. Meanwhile, $\Delta_{\text{dCE}}(f) \asymp \frac{1}{N}$.

Appendix B. Decision-theoretic guarantees: proofs and extensions

B.1. Proof of Proposition 5: ECE upper bounds the risk gap for binary decision loss

Proof [Lemma 6] First observe that

$$\begin{aligned}
&R_{\text{bd}}(f; \tau) - R_{\text{bd}}(h \circ f; \tau) \\
&= \tau\mathbb{E}[(1 - Y)(\mathbf{1}_{\{f(X)\geq\tau\}} - \mathbf{1}_{\{h\circ f(X)\geq\tau\}})] + (1 - \tau)\mathbb{E}[Y(\mathbf{1}_{\{f(X)<\tau\}} - \mathbf{1}_{\{h\circ f(X)<\tau\}})].
\end{aligned}$$

Then, turning differences of indicators into products of indicators, we get

$$\begin{aligned}
R_{\text{bd}}(f; \tau) - R_{\text{bd}}(h \circ f; \tau) =& \tau\mathbb{E}[(1 - Y)\mathbf{1}_{\{f(X)\geq\tau\}}\mathbf{1}_{\{h\circ f(X)<\tau\}}] \\
&- \tau\mathbb{E}[(1 - Y)\mathbf{1}_{\{f(X)<\tau\}}\mathbf{1}_{\{h\circ f(X)\geq\tau\}}] \\
&+ (1 - \tau)\mathbb{E}[Y\mathbf{1}_{\{f(X)<\tau\}}\mathbf{1}_{\{h\circ f(X)\geq\tau\}}] \\
&- (1 - \tau)\mathbb{E}[Y\mathbf{1}_{\{f(X)\geq\tau\}}\mathbf{1}_{\{h\circ f(X)<\tau\}}].
\end{aligned}$$

We achieve the desired result by distributing $(1-\tau)$ and $(1-Y)$ and collecting like terms. ■

Proof [Proposition 5] For any function h,

$$
\begin{aligned}
&R_{\mathrm{bd}}(f;\tau) - R_{\mathrm{bd}}(h\circ f;\tau)\\
&= \mathbb{E}[(\mathbb{E}[Y \mid f(X)] - \tau)\mathbf{1}_{\{f(X)<\tau\}}\mathbf{1}_{\{h\circ f(X)\geq\tau\}}] + \mathbb{E}[(\tau - \mathbb{E}[Y \mid f(X)])\mathbf{1}_{\{f(X)\geq\tau\}}\mathbf{1}_{\{h\circ f(X)<\tau\}}]\\
&\leq \mathbb{E}[(\mathbb{E}[Y \mid f(X)] - \tau)\mathbf{1}_{\{f(X)<\tau\}}\mathbf{1}_{\{\mathbb{E}[Y|f(X)]\geq\tau\}}] + \mathbb{E}[(\tau - \mathbb{E}[Y \mid f(X)])\mathbf{1}_{\{f(X)\geq\tau\}}\mathbf{1}_{\{\mathbb{E}[Y|f(X)]<\tau\}}]\\
&\leq \mathbb{E}[(\mathbb{E}[Y \mid f(X)] - f(X))\mathbf{1}_{\{f(X)<\tau\}}\mathbf{1}_{\{\mathbb{E}[Y|f(X)]\geq\tau\}}] + \mathbb{E}[(f(X) - \mathbb{E}[Y \mid f(X)])\mathbf{1}_{\{f(X)\geq\tau\}}\mathbf{1}_{\{\mathbb{E}[Y|f(X)]<\tau\}}]\\
&\leq \mathbb{E}[(\mathbb{E}[Y \mid f(X)] - f(X))\mathbf{1}_{\{f(X)<\mathbb{E}[Y|f(X)]\}}] + \mathbb{E}[(f(X) - \mathbb{E}[Y \mid f(X)])\mathbf{1}_{\{f(X)>\mathbb{E}[Y|f(X)]\}}]\\
&= \Delta_{\mathrm{ECE}}(f).
\end{aligned}
$$

■

B.2. Proof of Proposition 7: Cutoff Calibration Error upper bound the monotone risk gap for binary decision loss

Proof [Proposition 7] Let $h:[0,1]\to[0,1]$ be monotone.

Define $I, I' \subseteq [0,1]$ such that $\mathbf{1}_{\{h\circ f(X)\geq\tau\}} = \mathbf{1}_{\{f(X)\in I\}}$ and $\mathbf{1}_{\{h\circ f(X)<\tau\}} = \mathbf{1}_{\{f(X)\in I'\}}$. Since h is monotone, we know that I, I' must be intervals.

Then, using Lemma 6,

$$
\begin{aligned}
&R_{\mathrm{bd}}(f;\tau) - R_{\mathrm{bd}}(h\circ f;\tau)\\
&= \mathbb{E}[(Y-\tau)\mathbf{1}_{\{f(X)<\tau\}}\mathbf{1}_{\{h\circ f(X)\geq\tau\}}] + \mathbb{E}[(\tau - Y)\mathbf{1}_{\{f(X)\geq\tau\}}\mathbf{1}_{\{h\circ f(X)<\tau\}}]\\
&\leq \mathbb{E}[(Y-f(X))\mathbf{1}_{\{f(X)<\tau\}}\mathbf{1}_{\{h\circ f(X)\geq\tau\}}] + \mathbb{E}[(f(X) - Y)\mathbf{1}_{\{f(X)\geq\tau\}}\mathbf{1}_{\{h(f)<\tau\}}]\\
&= \mathbb{E}[(Y-f(X))\mathbf{1}_{\{f(X)<\tau\}}\mathbf{1}_{\{f(X)\in I\}}] + \mathbb{E}[(f(X) - Y)\mathbf{1}_{\{f(X)\geq\tau\}}\mathbf{1}_{\{f(X)\in I'\}}]\\
&\leq 2\Delta_{\mathrm{Cutoff}}(f).
\end{aligned}
$$

■

B.3. Cutoff Calibration Error upper bounds the monotone risk gap for bounded proper scoring rules, when $Y\in\{0,1\}$

We now generalize our results for ℓ_{bd} to a broader class of loss functions, still in the context of $Y\in\{0,1\}$. Like Hu and Wu (2024) and Kleinberg et al. (2023), we consider a loss function $\ell:\{0,1\}\times[0,1]\to\mathbb{R}$ that follows the very minimal constraint that it be a proper scoring rule:

$$
\mathbb{E}_{Y\sim\mathrm{Bernoulli}(p)}[\ell(Y,p)] \leq \mathbb{E}_{Y\sim\mathrm{Bernoulli}(q)}[\ell(Y,p)].
$$

While it may seem hopeless to be able to find a rearrangement like Lemma 6 that will aid us in relating the risk gap to a calibration measure for this minimally constrained ℓ, the fact that ℓ is a proper scoring rule proves to be enough. Loss functions that are proper scoring rules can be represented as mixtures across τ of $\ell_{\mathrm{bd}}(\cdot,\cdot;\tau)$: binary decision loss forms a basis for proper scoring rules. This representation dates back to Schervish (1989). Using a formulation of the Schervish representation from Gneiting and Raftery (2007), we get the following result.

Proposition 16 *Suppose $Y \in \{0,1\}$. Suppose $\ell : \{0,1\} \times [0,1] \to \mathbb{R}$ is a proper scoring rule that also satisfies the regularity conditions listed in Gneiting and Raftery (2007, Theorem 3). If the measure ν used to define the Schervish representation is a probability measure, then*

$$\mathbb{E}[\ell(Y, f(X))] - \inf_h \mathbb{E}[\ell(Y, h \circ f(X))] \leq \Delta_{\text{ECE}}(f)$$

and

$$\mathbb{E}[\ell(Y, f(X))] - \inf_{h\text{: monotone}} \mathbb{E}[\ell(Y, h \circ f(X))] \leq 2\Delta_{\text{Cutoff}}(f).$$

Proof Using the Schervish representation, we have

$$\mathbb{E}[\ell(Y, f(X))] - \mathbb{E}[\ell(Y, h \circ f(X))] = \mathbb{E}\left[\mathbb{E}_{\tau\sim\nu}[\ell_{\text{bd}}(Y, \mathbf{1}_{\{f(X)\geq\tau\}}, \tau) - \ell_{\text{bd}}(Y, \mathbf{1}_{\{h\circ f(X)\geq\tau\}}, \tau)]\right]$$

By Fubini's theorem,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_{\tau\sim\nu}[\ell_{\text{bd}}(Y, \mathbf{1}_{\{f(X)\geq\tau\}}, \tau) - \ell_{\text{bd}}(Y, \mathbf{1}_{\{h\circ f(X)\geq\tau\}}, \tau)]\right] \\
&= \mathbb{E}_{\tau\sim\nu}\left[\mathbb{E}[\ell_{\text{bd}}(Y, \mathbf{1}_{\{f(X)\geq\tau\}}, \tau) - \ell_{\text{bd}}(Y, \mathbf{1}_{\{h\circ f(X)\geq\tau\}}, \tau)]\right] \\
&\leq \sup_{\tau\in[0,1]} \mathbb{E}[\ell_{\text{bd}}(Y, \mathbf{1}_{\{f(X)\geq\tau\}}, \tau) - \ell_{\text{bd}}(Y, \mathbf{1}_{\{h\circ f(X)\geq\tau\}}, \tau)]
\end{aligned}$$

We conclude by applying Proposition 5 and Proposition 7. ■

B.4. Cutoff Calibration Error upper bounds the monotone risk gap of a sign-testing loss

When $Y \in [0,1]$, one may operate under the following sign-testing risk function (the "st" subscript refers to "sign testing"), with two actions and a pre-specified threshold Y^*,

$$R_{\text{st}}(f; Y^*) = \mathbb{E}[(Y - Y^*)\mathbf{1}_{\{f(X)\leq Y^*\}}\mathbf{1}_{\{Y>Y^*\}}] + \mathbb{E}[(Y^* - Y)\mathbf{1}_{\{f(X)>Y^*\}}\mathbf{1}_{\{Y\leq Y^*\}}].$$

Intuitively, we are trying to predict the sign of $Y - Y^*$, and the price we pay when we guess incorrectly is $|Y - Y^*|$.

Then, after rearrangement, we once again get, for arbitrary $h : [0,1] \to [0,1]$,

$$\begin{aligned}
R_{\text{st}}(f; Y^*) - R_{\text{st}}(h \circ f; Y^*) =&\mathbb{E}[(Y - Y^*)\mathbf{1}_{\{f(X)\leq Y^*\}}\mathbf{1}_{\{h\circ f(X)>Y^*\}}] \\
&+ \mathbb{E}[(Y^* - Y)\mathbf{1}_{\{f(X)>Y^*\}}\mathbf{1}_{\{h\circ f(X)\leq Y^*\}}].
\end{aligned}$$

We can use this identity to recover the same theoretical guarantees, in terms of Cutoff Calibration Error and ECE, as in the original binary decision framework.

Proposition 17 *For any $Y^* \in [0,1]$,*

$$R_{\text{st}}(f; Y^*) - \inf_{h \text{ monotone}} R_{\text{st}}(h \circ f; Y^*) \leq 2\Delta_{\text{Cutoff}}(f)$$

and

$$R_{\text{st}}(f; Y^*) - \inf_h R_{\text{st}}(h \circ f; Y^*) \leq \Delta_{\text{ECE}}(f).$$

The proof follows using the exact steps as those used to prove Proposition 5 and Proposition 7.

The fact that we recover the same theoretical guarantees as before is somewhat intuitive, given that the Bayes decision rule here has the same form as before: $\mathbf{1}_{\{\mathbb{E}[Y|f]\geq Y^*\}}$.

B.5. Implementation details of the experiment

In this section, we give details for the implementation of the experiment in Section 3.3.

In addition to the samples we used to fit f, we produce $N = 1000$ independent samples of (X, Y) pairs.

In all of the following procedures to approximate population quantities, we use $\mathbb{E}[Y \mid X]$ in place of $\mathbb{E}[Y \mid f(X)]$ since f is injective and X is a continuous random variable. Note that our procedures are sample efficient since we are plugging in $\mathbb{E}[Y \mid f(X)]$ in place of Y.

We approximate $\Delta_{\mathrm{dCE}}(f)$ using

$$\sup_{w \in \mathcal{L}_1} \frac{1}{N} \sum_{j=1}^{N} w(f(X_j))(\mathbb{E}[Y_j \mid X_j] - f(X_j)).$$

We approximate $\Delta_{\mathrm{Cutoff}}(f)$ using

$$\sup_{I\text{: interval}} \frac{1}{N} \sum_{j=1}^{N} (\mathbb{E}[Y_j \mid X_j] - f(X_j))\mathbf{1}_{\{f(X_j) \in I\}}.$$

We approximate $\Delta_{\mathrm{ECE}}(f)$ using

$$\frac{1}{N} \sum_{j=1}^{N} |\mathbb{E}[Y_j \mid X_j] - f(X_j)|.$$

We approximate $R_{\mathrm{bd}}(f;\tau)$ using

$$\frac{1}{N} \sum_{j=1}^{N} \ell_{\mathrm{bd}}(\mathbb{E}[Y_j \mid X_j], \mathbf{1}_{\{f(X_j) \geq \tau\}}; \tau).$$

We approximate $\inf_h R_{\mathrm{bd}}(h \circ f; \tau)$ using

$$\frac{1}{N} \sum_{j=1}^{N} \ell_{\mathrm{bd}}(\mathbb{E}[Y_j \mid X_j], \mathbf{1}_{\{\mathbb{E}[Y_j \mid X_j] \geq \tau\}}; \tau).$$

To approximate $\inf_{h\text{: monotone}} R_{\mathrm{bd}}(h \circ f; \tau)$, we minimize over all monotone decision rules:

$$\min\Bigg(\min_{\tau' \in \{0, f(X_1), \ldots, f(X_N), 1\}} \frac{1}{N} \sum_{j=1}^{N} \ell_{\mathrm{bd}}(\mathbb{E}[Y_j \mid X_j], \mathbf{1}_{\{f(X_j) \geq \tau'\}}; \tau),$$
$$\min_{\tau' \in \{0, f(X_1), \ldots, f(X_N), 1\}} \frac{1}{N} \sum_{j=1}^{N} \ell_{\mathrm{bd}}(\mathbb{E}[Y_j \mid X_j], \mathbf{1}_{\{f(X_j) \leq \tau'\}}; \tau)\Bigg).$$

Appendix C. Testability: proofs

C.1. Proof of Theorem 10: constraints on the effective support size of algorithms that strongly asymptotically control ECE

First we need a lemma, which we will prove in the end. Here, and for the remainder of the proof of this theorem, we will always write $\Delta_{\mathrm{ECE}}^{P}$ rather than Δ_{ECE} to clarify which distribution P is being used for computing the ECE.

Lemma 18 *Fix any $N > n \geq 1$. Let*

$$\mathcal{A} : (\mathcal{X} \times [0,1])^n \to \{ \textit{ measurable functions } f : \mathcal{X} \to [0,1]\}$$

be any procedure that inputs a data set $(X_1, Y_1), \ldots, (X_n, Y_n)$, and returns a fitted function $f_n = \mathcal{A}\big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$. Let $(X_1, Y_1), \ldots, (X_N, Y_N) \overset{iid}{\sim} P$, let $f_n = \mathcal{A}\big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$ be trained on the first n data points, and let $\widehat{P}_{N-n} = \frac{1}{N-n}\sum_{i=n+1}^{N} \delta_{(X_i,Y_i)}$ be the empirical distribution of the remaining $N - n$ data points. then

$$\mathbb{E}_{P^N}[\Delta_{\text{ECE}}^{\widehat{P}_{N-n}}(f_n)] \leq \frac{\sup_Q \mathbb{E}_{Q^n}[\Delta_{\text{ECE}}^{Q}(f_n)]}{1 - \frac{n(n-1)}{2N}} + \frac{2n}{N}.$$

Under the assumptions of Theorem 10, since $\Delta_{\text{ECE}} \leq 1$ always, we must have

$$\sup_Q \mathbb{E}_{Q^n}[\Delta_{\text{ECE}}^{Q}(f_n)] \leq c + \sup_Q \mathbb{P}_{Q^n}(\Delta_{\text{ECE}}^{Q}(f_n) > c) \leq c + \delta,$$

and consequently,

$$\mathbb{E}_{P^N}[\Delta_{\text{ECE}}^{\widehat{P}_{N-n}}(f_n)] \leq \frac{c + \delta}{1 - \frac{n(n-1)}{2N}} + \frac{2n}{N}.$$

Our next step is to work with this ECE error for the empirical distribution $\widehat{P}_{N-n}$. First consider a fixed function f. If for some $i \in \{n+1, \ldots, N\}$, $f(X_i)$ is unique among the values $f(X_{n+1}), \ldots, f(X_N)$, then $\mathbb{E}_{\widehat{P}_{N-n}}[Y \mid f(X) = f(X_i)] = Y_i$. In particular, this implies

$$\Delta_{\text{ECE}}^{\widehat{P}_{N-n}}(f) \geq \frac{1}{N-n} \sum_{i=n+1}^{N} |Y_i - f(X_i)| \cdot \mathbf{1}_{\{f(X_i) \neq f(X_j),\ \forall j \in \{n+1,\ldots,N\}\setminus\{i\}\}}.$$

Therefore, for a fixed function f,

$$\mathbb{E}\left[\Delta_{\text{ECE}}^{\widehat{P}_{N-n}}(f)\right] = \frac{1}{N-n} \sum_{i=n+1}^{N} \mathbb{E}\left[|Y_i - f(X_i)| \cdot \mathbf{1}_{\{f(X_i) \neq f(X_j),\ \forall j \in \{n+1,\ldots,N\}\setminus\{i\}\}}\right],$$

or equivalently by symmetry,

$$\mathbb{E}\left[\Delta_{\text{ECE}}^{\widehat{P}_{N-n}}(f)\right] = \mathbb{E}\left[|Y_{n+1} - f(X_{n+1})| \cdot \mathbf{1}_{\{f(X_{n+1}) \neq f(X_j),\ \forall j \in \{n+2,\ldots,N\}\}}\right].$$

Now define $p(x) = \mathbb{P}_P(f(X) = f(x))$. Then

$$\mathbb{E}\left[\mathbf{1}_{\{f(X_{n+1}) \neq f(X_j),\ \forall j \in \{n+2,\ldots,N\}\}} \mid X_{n+1}\right] = (1 - p(X_{n+1}))^{N-n-1} \geq e^{-1} \cdot \mathbf{1}_{\{p(X_{n+1}) \leq \frac{1}{N-n}\}},$$

since $(1 - \frac{1}{m})^{m-1} \geq e^{-1}$ for all $m \geq 1$, and therefore,

$$\mathbb{E}\left[\Delta_{\text{ECE}}^{\widehat{P}_{N-n}}(f)\right] \geq e^{-1} \cdot \mathbb{E}\left[|Y_{n+1} - f(X_{n+1})| \cdot \mathbf{1}_{\{p(X_{n+1}) \leq \frac{1}{N-n}\}}\right].$$

Since $Y \in [0,1]$, for any x, we must have

$$\begin{aligned}\mathrm{Var}(Y \mid X = x) &= \mathbb{E}[|Y - \mathbb{E}[Y \mid X]|^2 \mid X = x] \\ &= \inf_{t\in[0,1]} \mathbb{E}[|Y-t|^2 \mid X = x] \leq \inf_{t\in[0,1]} \mathbb{E}[|Y-t| \mid X = x],\end{aligned}$$

and so

$$\mathbb{E}\left[\Delta_{\mathrm{ECE}}^{\widehat{P}_{N-n}}(f)\right] \geq e^{-1} \cdot \mathbb{E}\left[\mathrm{Var}(Y_{n+1} \mid X_{n+1}) \cdot \mathbf{1}_{\{p(X_{n+1}) \leq \frac{1}{N-n}\}}\right].$$

Note that we must have $< N-n$ many values $t \in [0,1]$ such that $\mathbb{P}_P(f(X) = t) > \frac{1}{N-n}$, so

$$\mathbb{P}_P\Big(p(X) \leq \frac{1}{N-n}\Big) \geq 1 - \sup_{t_1,\dots,t_{N-n-1}} \mathbb{P}_P(f(X) \in \{t_1, \dots, t_{N-n-1}\}).$$

If $S_\gamma(f,P) \geq N-n$, then, we must have $\sup_{t_1,\dots,t_{N-n-1}} \mathbb{P}_P(f(X) \in \{t_1,\dots,t_{N-n-1}\}) < 1-\gamma$ and so $\mathbb{P}_P(p(X) \leq \frac{1}{N-n}) \geq \gamma$. Therefore,

$$\text{If } S_\gamma(f,P) \geq N-n, \text{ then } \mathbb{E}\left[\mathrm{Var}(Y_{n+1}|X_{n+1}) \cdot \mathbf{1}_{\{p(X_{n+1}) \leq \frac{1}{N-n}\}}\right] \geq \sigma^2_\gamma(P),$$

by definition of $\sigma^2_\gamma(P)$. In other words,

$$\text{If } S_\gamma(f,P) \geq N-n, \text{ then } \mathbb{E}\left[\Delta_{\mathrm{ECE}}^{\widehat{P}_{N-n}}(f)\right] \geq e^{-1}\sigma^2_\gamma(P).$$

Now we plug in $f = f_n$, and condition on f_n:

$$\text{If } S_\gamma(f_n,P) \geq N-n, \text{ then } \mathbb{E}\left[\Delta_{\mathrm{ECE}}^{\widehat{P}_{N-n}}(f_n) \mid f_n\right] \geq e^{-1}\sigma^2_\gamma(P).$$

(Here, since we are conditioning on f_n, the expected value is being taken with respect to the remaining random sample, $(X_{n+1}, Y_{n+1}), \dots, (X_N, Y_N)$.) Finally, marginalizing over f_n,

$$\mathbb{E}\left[\Delta_{\mathrm{ECE}}^{\widehat{P}_{N-n}}(f_n)\right] \geq e^{-1}\sigma^2_\gamma(P) \cdot \mathbb{P}(S_\gamma(f_n,P) \geq N-n).$$

Returning to the results of the lemma, then,

$$e^{-1}\sigma^2_\gamma(P) \cdot \mathbb{P}(S_\gamma(f_n,P) \geq N-n) \leq \frac{c+\delta}{1 - \frac{n(n-1)}{2N}} + \frac{2n}{N}.$$

Finally, choosing $N = n^2$, we obtain

$$e^{-1}\sigma^2_\gamma(P) \cdot \mathbb{P}(S_\gamma(f_n,P) \geq n^2) \leq \frac{c+\delta}{1 - \frac{n(n-1)}{2n^2}} + \frac{2}{n} \leq 2(c+\delta+n^{-1}),$$

which completes the proof.

Finally, we prove the lemma.

Proof [Lemma 18] First, let $\widehat{P}_N = \frac{1}{N}\sum_{i=1}^N \delta_{(X_i,Y_i)}$ be the empirical distribution of the full sample of size N. For the moment, treat $(X_1,Y_1),\dots,(X_N,Y_N)$ as fixed. Then

$$\mathbb{E}_{(\widehat{P}_N)^n}[\Delta_{\mathrm{ECE}}^{\widehat{P}_N}(f_n)] \le \sup_Q \mathbb{E}_{Q^n}[\Delta_{\mathrm{ECE}}^{Q}(f_n)]$$

since $Q = \widehat{P}_N$ is a valid distribution. To be more explicit, on the left-hand side, we have f_n being the output of $\mathcal{A}$ when trained on a sample of size n drawn i.i.d. from $\widehat{P}_N$, or equivalently,

$$f_n = \mathcal{A}((X_{i_1},Y_{i_1}),\dots,(X_{i_n},Y_{i_n}))$$

where $i_1,\dots,i_n \stackrel{iid}{\sim} \mathrm{Uniform}([N])$. Now, with probability $(1-\frac{1}{N})\cdot\ldots\cdot(1-\frac{n-1}{N}) \ge 1-\frac{n(n-1)}{2N}$, the indices $i_1,\dots,i_n$ are all distinct—i.e., equivalently, these indices have been sampled without replacement (WOR). Therefore,

$$\begin{aligned}&\mathbb{E}_{i_1,\dots,i_n\sim\text{WOR}}[\Delta_{\mathrm{ECE}}^{\widehat{P}_N}(\mathcal{A}((X_{i_1},Y_{i_1}),\dots,(X_{i_n},Y_{i_n})))]\\ &\qquad= \mathbb{E}_{(\widehat{P}_N)^n}[\Delta_{\mathrm{ECE}}^{\widehat{P}_N}(\mathcal{A}((X_{i_1},Y_{i_1}),\dots,(X_{i_n},Y_{i_n}))) \mid i_1,\dots,i_n \text{ distinct}]\\ &\qquad\le \frac{\mathbb{E}_{(\widehat{P}_N)^n}[\Delta_{\mathrm{ECE}}^{\widehat{P}_N}(\mathcal{A}((X_{i_1},Y_{i_1}),\dots,(X_{i_n},Y_{i_n})))\cdot\mathbf{1}_{\{i_1,\dots,i_n \text{ distinct}\}}]}{1-\frac{n(n-1)}{2N}}\\ &\qquad\qquad\le \frac{\mathbb{E}_{(\widehat{P}_N)^n}[\Delta_{\mathrm{ECE}}^{\widehat{P}_N}(\mathcal{A}((X_{i_1},Y_{i_1}),\dots,(X_{i_n},Y_{i_n})))]}{1-\frac{n(n-1)}{2N}},\end{aligned}$$

where the first expected value is taken with respect to $i_1,\dots,i_n$ sampled uniformly without replacement from $[N]$.

Now, take the expected value over $(X_1,Y_1),\dots,(X_N,Y_N) \stackrel{iid}{\sim} P$. We then have

$$\mathbb{E}\left[\Delta_{\mathrm{ECE}}^{\widehat{P}_N}(\mathcal{A}((X_{i_1},Y_{i_1}),\dots,(X_{i_n},Y_{i_n})))\right] \le \frac{\sup_Q \mathbb{E}_{Q^n}[\Delta_{\mathrm{ECE}}^{Q}(f_n)]}{1-\frac{n(n-1)}{2N}}$$

where, on the left-hand side, the expected value is taken with respect to both the draw of $(X_1,Y_1),\dots,$ $(X_N,Y_N) \stackrel{iid}{\sim} P$, and also $i_1,\dots,i_n$ sampled uniformly without replacement from $[N]$. By symmetry, the data points are invariant to permutation, so it's equivalent to write

$$\mathbb{E}_{P^N}\left[\Delta_{\mathrm{ECE}}^{\widehat{P}_N}(f_n)\right] \le \frac{\sup_Q \mathbb{E}_{Q^n}[\Delta_{\mathrm{ECE}}^{Q}(f_n)]}{1-\frac{n(n-1)}{2N}}$$

where now on the left-hand side, $f_n = \mathcal{A}((X_1,Y_1),\dots,(X_n,Y_n))$ is simply trained on the first n data points.

Finally, let P and Q be any distributions, and let $f : \mathcal{X} \to [0,1]$ be any function. Then for any function $w : [0,1] \to [-1,1]$,

$$\mathbb{E}_P[w(f(X))\cdot(Y-f(X))] \le \mathbb{E}_Q[w(f(X))\cdot(Y-f(X))] + 2d_{TV}(P,Q),$$

since $w(f(X)) \cdot (Y - f(X)) \in [-1, 1]$. So,

$$\begin{aligned}\Delta^P_{\mathrm{ECE}}(f) &= \sup_{w:[0,1]\to[-1,1]} \mathbb{E}_P[w(f(X)) \cdot (Y - f(X))] \\ &\leq \sup_{w:[0,1]\to[-1,1]} \mathbb{E}_Q[w(f(X)) \cdot (Y - f(X))] + 2d_{TV}(P, Q) \\ &= \Delta^Q_{\mathrm{ECE}}(f) + 2d_{TV}(P, Q).\end{aligned}$$

Consequently, it holds almost surely that

$$\Delta^{\widehat{P}_{N-n}}_{\mathrm{ECE}}(f_n) \leq \Delta^{\widehat{P}_N}_{\mathrm{ECE}}(f_n) + 2d_{TV}(\widehat{P}_{N-n}, \widehat{P}_N) \leq \Delta^{\widehat{P}_N}_{\mathrm{ECE}}(f_n) + \frac{2n}{N}.$$

This completes the proof. ■

C.2. Hardness of asymptotic ECE calibration

Our hardness result for ECE, stated in Theorem 10, is closely related to an asymptotic hardness result of Gupta et al. (2020). Here we compare between the two, and show how our result implies the existing result—in particular, ours gives a more precise characterization by providing a finite-sample bound on effective support size of any procedure that guarantees ECE calibration, but asymptotically can be interpreted in an analogous way.

First we define what it means for an algorithm to provide an asymptotic ECE guarantee.

Definition 19 *Consider an algorithm*

$$\mathcal{A} : \cup_{n\geq 0}(\mathcal{X} \times [0, 1])^n \to \{\text{ measurable functions } f : \mathcal{X} \to [0, 1]\}.$$

We say that ***an algorithm $\mathcal{A}$ is strongly asymptotically ECE-calibrated*** *if*

$$\limsup_{n\to\infty} \sup_P \mathbb{E}_{P^n}[\Delta_{\mathrm{ECE}}(f_n)] = 0,$$

where $\sup_P$ takes the supremum over all distributions P on $\mathcal{X} \times [0, 1]$ and where $f_n := \mathcal{A}(\mathcal{D}_n)$ for data set $\mathcal{D}_n := \{(X_i, Y_i)\}_{i\in[n]} \sim P^n$.

Gupta et al. (2020) show that any post-hoc calibration method returning an injective map on the pretrained model, cannot satisfy this property—that is, if $\mathcal{A}$ has the potential to produce a map whose output is continuously distributed on $[0, 1]$. We will now see how Theorem 10 implies this same takeaway message: the asymptotic calibration property is impossible to attain unless we bound the rate at which the support size of the output f_n can grow.

Corollary 20 *Suppose $\mathcal{A}$ is strongly asymptotically ECE-calibrated (in the sense of Definition 19). Then there exists a sequence $\nu_n \to 0$ such that*

$$\mathbb{P}_{P^n}\left\{S_\gamma(f_n, P) \leq n^2\right\} \geq 1 - \frac{\nu_n}{\sigma^2_\gamma(P)} \text{ for all distributions } P \text{ and all } \gamma.$$

We can interpret this as saying that any function f_n, returned by a post-hoc calibration procedure with asymptotic ECE guarantees, must necessarily be (mostly) discrete.

Proof Define

$$\epsilon_n = \sup_P \mathbb{E}_{P^n}[\Delta^P_{\text{ECE}}(f_n)].$$

If $\mathcal{A}$ is strongly asymptotically calibrated, then we must have

$$\epsilon_n \to 0.$$

Then

$$\sup_P \mathbb{E}_{P^n}[\Delta^P_{\text{ECE}}(f_n)] \le \epsilon_n \implies \mathbb{P}_{P^n}(\Delta^P_{\text{ECE}}(f_n) \le \epsilon_n^{1/2}) \ge 1 - \epsilon_n^{1/2} \text{ for all } P,$$

by Markov's inequality. Applying Theorem 10 with $c = \delta = \epsilon_n^{1/2}$, then, for all γ,

$$\mathbb{P}_{P^n}\left\{S_\gamma(f_n, P) \le n^2\right\} \ge 1 - \frac{2e(2\epsilon_n^{1/2} + n^{-1})}{\sigma^2_\gamma(P)} \text{ for all distributions } P.$$

Since $2e(2\epsilon_n^{1/2} + n^{-1}) \to 0$ this completes the proof. ■

C.3. Proof of Corollary 9

We will prove the stronger statement

$$\mathbb{P}\big\{\Delta_{\text{Cutoff}}(f_n) \le c \mid (X_1, Y_1), \dots, (X_{\lceil n/2\rceil}, Y_{\lceil n/2\rceil})\big\} \ge 1 - 2\delta.$$

In other words, we will condition on the first batch of data (used to train $f_{n,\text{init}}$), and the probability is computed with respect to the distribution the remaining half of the data, $(X_{\lceil n/2\rceil+1}, Y_{\lceil n/2\rceil+1}), \dots, (X_n, Y_n) \overset{iid}{\sim} P$.

We will define two events:

$$\mathcal{E}_1 = \left\{\left|\widehat{\Delta}_{\text{Cutoff}}(f_{n,\text{init}}) - \Delta^P_{\text{Cutoff}}(f_{n,\text{init}})\right| \le \frac{20 + \sqrt{2\log(1/\delta)}}{\sqrt{\lfloor n/2\rfloor}}\right\}$$

and

$$\mathcal{E}_2 = \left\{\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=\lceil n/2\rceil+1}^{n} Y_i - \mathbb{E}_P[Y]\right| \le \sqrt{\frac{\log(1/\delta)}{2\lfloor n/2\rfloor}}\right\}.$$

By Proposition 8 (applied to the data set $\{(X_i, Y_i)\}_{i=\lceil n/2\rceil+1}^n$, i.e., we have sample size $\lfloor n/2\rfloor$ in place of n), we have

$$\mathbb{P}\big\{\mathcal{E}_1 \mid (X_1, Y_1), \dots, (X_{\lceil n/2\rceil}, Y_{\lceil n/2\rceil})\big\} \ge 1 - \delta.$$

Moreover, by Hoeffding's inequality (again, applied with sample size $\lfloor n/2\rfloor$), we have

$$\mathbb{P}\big\{\mathcal{E}_2 \mid (X_1, Y_1), \dots, (X_{\lceil n/2\rceil}, Y_{\lceil n/2\rceil})\big\} \ge 1 - \delta.$$

To complete the proof, then, it suffices to show that on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, it must hold that f_n is calibrated. To see why this is true, we split into cases:

- If $\widehat{\Delta}_{\text{Cutoff}}(f_{n,\text{init}}) \leq c - \frac{20+\sqrt{2\log(1/\delta)}}{\sqrt{\lfloor n/2 \rfloor}}$, then our procedure defines $f_n = f_{n,\text{init}}$. And, we have

$$\Delta_{\text{Cutoff}}(f_n) = \Delta_{\text{Cutoff}}(f_{n,\text{init}}) \leq \widehat{\Delta}_{\text{Cutoff}}(f_{n,\text{init}}) + \frac{20+\sqrt{2\log(1/\delta)}}{\sqrt{\lfloor n/2 \rfloor}} \leq c,$$

  where the first inequality holds because we have assumed the event $\mathcal{E}_1$ holds.

- If instead $\widehat{\Delta}_{\text{Cutoff}}(f_{n,\text{init}}) > c - \frac{20+\sqrt{2\log(1/\delta)}}{\sqrt{\lfloor n/2 \rfloor}}$, then our procedure defines $f_n = f_{n,\text{const}}$. And, we have

$$\Delta_{\text{Cutoff}}(f_n) = \Delta_{\text{Cutoff}}(f_{n,\text{const}}) = \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=\lceil n/2 \rceil + 1}^{n} Y_i - \mathbb{E}_P[Y] \right| \leq \sqrt{\frac{\log(1/\delta)}{2\lfloor n/2 \rfloor}} \leq c,$$

  where the last inequality holds by assumption on c, the next-to-last inequality holds due to the event $\mathcal{E}_2$, and the second equality holds since, for any constant function $f(x) \equiv C$, its calibration error is given by $\Delta_{\text{Cutoff}}(f) = |\mathbb{E}[Y] - C|$.

We have therefore verified that $\Delta_{\text{Cutoff}}(f_n) \leq c$ in both cases, and so

$$\begin{aligned}\mathbb{P}\big\{\Delta_{\text{Cutoff}}(f_n) \leq c \mid (X_1, Y_1), \ldots, (X_{\lceil n/2 \rceil}, Y_{\lceil n/2 \rceil})\big\} \\ \geq \mathbb{P}\big\{\mathcal{E}_1 \cap \mathcal{E}_2 \mid (X_1, Y_1), \ldots, (X_{\lceil n/2 \rceil}, Y_{\lceil n/2 \rceil})\big\} \geq 1 - 2\delta.\end{aligned}$$

C.4. Proof of Proposition 8: Cutoff Calibration Error can be estimated at the parametric rate for fixed f

Proof [Proposition 8] We proceed by turning the problem into a form for which we can apply empirical process theory. For notational brevity, we adopt the notation that $\mathbb{E}_n[g(X,Y)] := \frac{1}{n}\sum_{i=1}^n g(X_i, Y_i)$, where g is an arbitrary function.

Observe that

$$\begin{aligned}&|\widehat{\Delta}_{\text{Cutoff}}(f) - \Delta_{\text{Cutoff}}(f)| \\ &= \left| \sup_{I:\text{ interval}} \left|\mathbb{E}\left[(Y - f(X))\mathbf{1}_{\{f(X) \in I\}}\right]\right| - \sup_{I':\text{ interval}} \left|\mathbb{E}_n\left[(Y - f(X))\mathbf{1}_{\{f(X) \in I'\}}\right]\right| \right| \\ &\leq \sup_{I:\text{ interval}} ||\mathbb{E}[(Y - f(X))\mathbf{1}_{\{f(X) \in I\}}]| - |\mathbb{E}_n[(Y - f(X))\mathbf{1}_{\{f(X) \in I\}}]|| \\ &\leq \sup_{I:\text{ interval}} |\mathbb{E}[(Y - f(X))\mathbf{1}_{\{f(X) \in I\}}] - \mathbb{E}_n[(Y - f(X))\mathbf{1}_{\{f(X) \in I\}}]|,\end{aligned}$$

where the first inequality follows from the triangle inequality of suprema.

We can now apply empirical process theory. By Koltchinskii (2011, Theorem 2.2), we know that the relevant Rademacher complexity of this empirical process is bounded by twice the Rademacher complexity of

$$\{g : [0,1] \to \{0,1\} \mid g(z) = \mathbf{1}_{\{x \in I\}} \text{ for some } I \subseteq [0,1] \text{ interval}\}. \tag{4}$$

By Lemma 21, we know that the Rademacher complexity of (4) is at most $5/\sqrt{n}$.

Therefore, by Wainwright (2019, Theorem 4.10), we know that $\forall \epsilon \geq 0$

$$\mathbb{P}\{|\widehat{\Delta}_{\text{Cutoff}}(f) - \Delta_{\text{Cutoff}}(f)| \leq 20/\sqrt{n} + \epsilon\} \geq 1 - \exp(-n\epsilon^2/2).$$

Thus, $\forall \delta > 0$,

$$|\widehat{\Delta}_{\text{Cutoff}}(f) - \Delta_{\text{Cutoff}}(f)| \leq \frac{20 + \sqrt{2\log(1/\delta)}}{\sqrt{n}}$$

with probability at least $1 - \delta$. ■

Lemma 21 *Let $\epsilon_i, \ldots, \epsilon_n$ be iid Rademacher random variables. Then,*

$$\mathbb{E}\left[\sup_{\substack{f\in[-1,1]^n \\ \sum_{i=1}^{n=1} |f_{i+1}-f_i| \leq M}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f_i\right] \leq \frac{2M+1}{\sqrt{n}}.$$

Proof Observe that

$$\begin{aligned}\sum_{i=1}^{n} \epsilon_i f_i &= f_1 \sum_{i=1}^{n} \epsilon_i + \sum_{i=2}^{n} \epsilon_i (f_i - f_1) \\ &= f_1 \sum_{i=1}^{n} \epsilon_i + \sum_{i=2}^{n} \epsilon_i \sum_{j=1}^{i-1} (f_{j+1} - f_j) \\ &= f_1 \sum_{i=1}^{n} \epsilon_i + \sum_{j=1}^{n-1} (f_{j+1} - f_j) \sum_{i=j+1}^{n} \epsilon_i.\end{aligned}$$

Therefore,

$$\sup_{\substack{f\in[-1,1]^n \\ \sum_{i=1}^{n=1} |f_{i+1}-f_i| \leq M}} \sum_{i=1}^{n} \epsilon_i f_i \leq \left|\sum_{i=1}^{n} \epsilon_i\right| + M \max_{1\leq j\leq n-1} \left|\sum_{i=j+1}^{n} \epsilon_i\right|.$$

Using an Jensen's inequality and an easy-to-show fact of random walks (see, e.g., Lawler (2010)), we get

$$\mathbb{E}\left|\sum_{i=1}^{n} \epsilon_i\right| \leq \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{n} \epsilon_i\right)^2\right]} = \sqrt{n}.$$

Using the Lévy inequality, we similarly get

$$\begin{aligned}
\mathbb{E}\left[\max_{1\le j\le n-1}\left|\sum_{i=j+1}^{n}\epsilon_i\right|\right] &= \mathbb{E}\left[\max_{1\le j\le n-1}\left|\sum_{i=1}^{j}\epsilon_i\right|\right] \\
&= \int_0^\infty \mathbb{P}\left\{\max_{1\le j\le n-1}\left|\sum_{i=1}^{j}\epsilon_i\right| \ge t\right\} dt \\
&\le 2\int_0^\infty \mathbb{P}\left\{\left|\sum_{i=1}^{n-1}\epsilon_i\right| \ge t\right\} dt \\
&= 2\mathbb{E}\left|\sum_{i=1}^{n-1}\epsilon_i\right| \\
&= 2\sqrt{n-1}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\sup_{\substack{f\in[-1,1]^n \\ \sum_{i=1}^{n=1}|f_{i+1}-f_i|\le M}} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_i\right] \le \frac{\sqrt{n}+2M\sqrt{n-1}}{n} \le \frac{2M+1}{\sqrt{n}}.$$

■

Appendix D. Post-hoc calibration: proofs and counter-examples

D.1. Proof of Proposition 11: isotonic regression enjoys uniform asymptotic control of Cutoff Calibration Error

Proof [Proposition 11] Observe that

$$\begin{aligned}
\Delta^P_{\text{Cutoff}}(h_n \circ f) \le \sup_{I\text{: interval}} &|\mathbb{E}[Y\mathbf{1}_{\{h_n(f(X))\in I\}} \mid h_n] - \mathbb{E}_n[h_n(f(X))\mathbf{1}_{\{h_n(f(X))\in I\}}]| \\
&+ \sup_{I\text{: interval}} |\mathbb{E}_n[h_n(f(X))\mathbf{1}_{\{h_n(f(X))\in I\}}] - \mathbb{E}[h_n(f(X))\mathbf{1}_{\{h_n(f(X))\in I\}} \mid h_n]|.
\end{aligned}$$

Then, using Lemma 22,

$$\begin{aligned}
\Delta^P_{\text{Cutoff}}(h_n \circ f) \le \sup_{I\text{: interval}} &|\mathbb{E}[Y\mathbf{1}_{\{h_n(f(X))\in I\}} \mid h_n] - \mathbb{E}_n[Y\mathbf{1}_{\{h_n(f(X))\in I\}}]| \qquad (5) \\
&+ \sup_{I\text{: interval}} |\mathbb{E}_n[h_n(f(X))\mathbf{1}_{\{h_n(f(X))\in I\}}] - \mathbb{E}[h_n(f(X))\mathbf{1}_{\{h_n(f(X))\in I\}} \mid h_n]|.
\end{aligned}$$

Since h_n is monotone, we know that if $I \subseteq [0,1]$ is an interval, then $h_n^{-1}(I)$ must also be an interval. Therefore,

$$\begin{aligned}
\Delta^P_{\text{Cutoff}}(h_n \circ f) \le \sup_{I\text{: interval}} &|\mathbb{E}[Y\mathbf{1}_{\{f(X)\in I\}} \mid h_n] - \mathbb{E}_n[Y\mathbf{1}_{\{f(X)\in I\}}]| \\
&+ \sup_{I\text{: interval}} |\mathbb{E}_n[h_n(f(X))\mathbf{1}_{\{f(X)\in I\}}] - \mathbb{E}[h_n(f(X))\mathbf{1}_{\{f(X)\in I\}} \mid h_n]|.
\end{aligned}$$

Then, using Lemma 23, we get

$$\begin{aligned}\Delta^P_{\text{Cutoff}}(h_n \circ f) \leq & \sup_{I\text{: interval}} |\mathbb{E}[Y\mathbf{1}_{\{f(X)\in I\}} \mid h_n] - \mathbb{E}_n[Y\mathbf{1}_{\{f(X)\in I\}}]| \\ & + \sup_{I\text{: interval}} |\mathbb{E}_n[\mathbf{1}_{\{f(X)\in I\}}] - \mathbb{E}[\mathbf{1}_{\{f(X)\in I\}} \mid h_n]|.\end{aligned}$$

Both terms no longer depends on h_n, so we can rewrite this as

$$\begin{aligned}\Delta^P_{\text{Cutoff}}(h_n \circ f) \leq & \sup_{I\text{: interval}} |\mathbb{E}[Y\mathbf{1}_{\{f(X)\in I\}}] - \mathbb{E}_n[Y\mathbf{1}_{\{f(X)\in I\}}]| \\ & + \sup_{I\text{: interval}} |\mathbb{E}_n[\mathbf{1}_{\{f(X)\in I\}}] - \mathbb{E}[\mathbf{1}_{\{f(X)\in I\}}]|.\end{aligned}$$

Using a similar Rademacher complexity argument from Proposition 8, we know with probability at least $1 - \delta/2$ that each term individually is bounded above by $\frac{20+\sqrt{2\log(2/\delta)}}{\sqrt{n}}$ and $\frac{10+\sqrt{2\log(2/\delta)}}{\sqrt{n}}$, respectively. Note that the high-probability bound for the first term is higher than the second due to the contraction mapping argument doubling the Rademacher complexity.

Therefore, by the union bound, we know that

$$\Delta^P_{\text{Cutoff}}(h_n \circ f) \leq \frac{30 + 2\sqrt{2\log(2/\delta)}}{\sqrt{n}},$$

with probability at least $1 - \delta$. ■

Lemma 22 *Let $\hat{f}$ refer to running isotonic regression on $\{(X_i, Y_i)\}_{i\in[n]}$, where $Y \in \mathbb{R}$. Then, for any $A \subseteq \mathbb{R}$,*

$$\frac{1}{n}\sum_{i=1}^n \hat{f}(X_i)\mathbf{1}_{\{\hat{f}(X_i)\in A\}} = \frac{1}{n}\sum_{i=1}^n Y_i\mathbf{1}_{\{\hat{f}(X_i)\in A\}}.$$

Proof This follows immediately from the properties of isotonic regression: namely, it is a piecewise empirical mean. ■

Lemma 23 *For any unimodal $g : \mathbb{R} \to [0, 1]$ and any distributions P, Q on $\mathbb{R}$,*

$$|\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)]| \leq \sup_{\textit{Interval } I\subseteq\mathbb{R}} |P(I) - Q(I)|.$$

Proof Observe that

$$\begin{aligned}|\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)]| &= \left|\int_{\mathbb{R}} g(x)(dP - dQ)(x)\right| \\ &= \left|\int_{\mathbb{R}}\int_0^1 \mathbf{1}_{\{t\leq g(x)\}}dt(dP - dQ)(x)\right| \\ &= \left|\int_0^1\int_{\mathbb{R}} \mathbf{1}_{\{g(x)\geq t\}}(dP - dQ)(x)dt\right| \\ &\leq \sup_{t\in[0,1]}\left|\int_{\mathbb{R}} \mathbf{1}_{\{g(x)\geq t\}}(dP - dQ)(x)\right|.\end{aligned}$$

Let

$$I := \{x : g(x) \geq t\}.$$

Since g is unimodal, we know that I must be an interval.

Therefore,

$$\sup_{t \in [0,1]} \left| \int_{\mathbb{R}} \mathbf{1}_{\{g(x) \geq t\}} (dP - dQ)(x) \right| \leq \sup_{\text{Interval } I \subseteq \mathbb{R}} |P(I) - Q(I)|.$$

■

D.2. Modifying Platt scaling to achieve Cutoff calibration

In this section, we propose a modified Platt scaling procedure that is able to achieve asymptotic control of Cutoff Calibration Error by testing for model misspecification. This may be beneficial to users who prefer Platt scaling due to its empirical success in the limited sample regime (Niculescu-Mizil and Caruana, 2005).

Define the usual Platt-scaling post-processor as $\tilde{h}_n$. Then, our final h_n is designed to ensure $\widehat{\Delta}_{\text{Cutoff}}(h_n \circ f)$ is small, up to a user-specified threshold ϵ_n:

$$h_n(z) = \begin{cases} \tilde{h}_n(z) & \text{if } \widehat{\Delta}_{\text{Cutoff}}(\tilde{h}_n \circ f) \leq \epsilon_n \\ \mathbb{E}_n[Y] & \text{otherwise} \end{cases}.$$

We can ensure that this procedure achieves a similar asymptotic guarantee to our isotonic regression result.

Proposition 24 *Suppose* $(X_i, Y_i) \overset{iid}{\sim} P$ *for* $i = 1, \ldots, n$. *Assume* $Y_i \in [0,1]$ *and* f *is fixed. Let* h_n *be as defined in* (D.2). *Fix* $\delta > 0$. *Let* $\epsilon_n > 0$ *be a user-specified tolerance for model-misspecification. Then, with probability at least* $1 - \delta$,

$$\Delta^P_{\text{Cutoff}}(h_n \circ f) \leq \frac{30 + 2\sqrt{2\log(2/\delta)}}{\sqrt{n}} + \epsilon_n.$$

We note that it may be natural to set $\epsilon_n = \frac{20+\sqrt{2\log(20)}}{\sqrt{n}}$, which corresponds to the 95% upper confidence bound of $\widehat{\Delta}_{\text{Cutoff}}(f)$ when $\Delta_{\text{Cutoff}}(f) = 0$.

Proof Observe that returning the unconditional empirical average of Y ensures that $\widehat{\Delta}_{\text{Cutoff}}$ is 0:

$$\widehat{\Delta}_{\text{Cutoff}}(\mathbb{E}_n[Y]) = 0.$$

Therefore, we know

$$\widehat{\Delta}_{\text{Cutoff}}(h_n \circ f) \leq \epsilon_n.$$

Then,

$$\begin{aligned} &\Delta^P_{\text{Cutoff}}(h_n \circ f) \\ &= \Delta^P_{\text{Cutoff}}(h_n \circ f) - \widehat{\Delta}_{\text{Cutoff}}(h_n \circ f) + \widehat{\Delta}_{\text{Cutoff}}(h_n \circ f) \\ &\leq \Delta^P_{\text{Cutoff}}(h_n \circ f) - \widehat{\Delta}_{\text{Cutoff}}(h_n \circ f) + \epsilon_n \\ &\leq \sup_{I\text{: interval}} \left| \mathbb{E}[(Y - h_n \circ f(X))\mathbf{1}_{\{h_n \circ f(X) \in I\}}] - \frac{1}{n}\sum_{i=1}^{n}(Y_i - h_n \circ f(X_i))\mathbf{1}_{\{h_n \circ f(X_i) \in I\}} \right| + \epsilon_n. \end{aligned}$$

The first term can be shown to be $\lesssim \frac{1}{\sqrt{n}}$ with probability at least $1-\delta$ following the same arguments in the isotonic regression proof, from equation (5) onward, since h_n is also monotone here. ■