The Importance of Being Smoothly Calibrated

Parikshit Gopalan
Apple, Palo Alto, CA, USA

Konstantinos Stavropoulos[1]
University of Texas at Austin, TX, USA

Kunal Talwar
Apple, Palo Alto, CA, USA

Pranay Tankala[2]
Harvard University, Cambridge, MA, USA

Abstract

Recent work has highlighted the centrality of smooth calibration [12] as a robust measure of calibration error. We generalize, unify, and extend previous results on smooth calibration, both as a robust calibration measure, and as a step towards omniprediction, which enables predictions with low regret for downstream decision makers seeking to optimize some proper loss unknown to the predictor.

- We present a new omniprediction guarantee for smoothly calibrated predictors, for the class of all bounded proper losses. We smooth the predictor by adding some noise to it, and compete against smoothed versions of any benchmark predictor on the space, where we add some noise to the predictor and then post-process it arbitrarily. The omniprediction error is bounded by the smooth calibration error of the predictor and the earth mover's distance from the benchmark. We exhibit instances showing that this dependence cannot, in general, be improved. We show how this unifies and extends prior results [5, 10] on omniprediction from smooth calibration.
- We present a crisp new characterization of smooth calibration in terms of the earth mover's distance to the closest perfectly calibrated joint distribution of predictions and labels. This also yields a simpler proof of the relation to the *lower distance to calibration* from [1].
- We use this to show that the *upper distance to calibration* cannot be estimated within a quadratic factor with sample complexity independent of the support size of the predictions. This is in contrast to the *distance to calibration*, where the corresponding problem was known to be information-theoretically impossible: no finite number of samples suffice [1].

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases Smooth Calibration, Omniprediction, Distance to Calibration

Digital Object Identifier 10.4230/LIPIcs.FORC.2026.21

Related Version *Full Version*: https://arxiv.org/abs/2603.16015

1 Introduction

Consider the setting of binary classification, where we wish to learn a predictor $p : \mathcal{X} \to [0, 1]$ based on labeled samples $(\mathbf{x}, \mathbf{y})$ drawn from a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$. The prediction $p(x)$ represents our estimate of the conditional probability at x that the label is 1. Perfect *calibration*, a classical notion which originates in the forecasting literature [4], requires that $\mathbf{E}[\mathbf{y}|p(\mathbf{x})] = p(\mathbf{x})$. Calibration guarantees several desirable properties to any downstream decision maker who uses these predictions to minimize a *proper* loss function. These guarantees have the flavor of ensuring that the decision maker can *trust* the predictor, as if it were the Bayes optimal predictor.

[1] Work done during an internship at Apple.
[2] Work done during an internship at Apple.

The property of calibration that is most relevant to our work is that it ensures no-regret with respect to all post-processings; this was first shown in the classic work of Foster and Vohra [5]. Indeed, a predictor p is calibrated if and only if p has lower expected loss than $\kappa \circ p$ for all post-processing functions $\kappa : [0,1] \to [0,1]$ and all proper loss functions $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$. Restated in the language of omniprediction (introduced in [8] and applied in the context of calibration by [10, 9]), a calibrated predictor is a perfect $(\mathcal{L}, \mathcal{C})$-omnipredictor, where the loss family $\mathcal{L}$ comprises all proper loss functions, and the hypothesis class $\mathcal{C}$ comprises all post-processings of p itself.

Perfect calibration is not an achievable goal in practice for both computational and information-theoretic reasons. This has led to a long line of work that aims at finding notions of approximate calibration that are more computationally tractable to achieve and to verify, and which preserve the desirable properties of calibration for decision making. Much of this work focuses on two main questions:

- How should we measure the calibration error of our predictor, so as to have a measure that is both robust and efficient? This question has been studied in [12, 14, 1, 2, 3].
- What kind of loss minimization guarantees for downstream decision makers can we get from approximate notions of calibration? This question has been studied in [2, 11, 9].

We refer the reader to the recent survey [6] for more details.

The work of [1] suggested a property testing-inspired approach to the first question: measure the calibration error of predictor by the distance from the nearest perfectly calibrated predictor. They refer to this notion as the *distance to calibration*. However, implementing this approach immediately runs into some basic questions: How do we measure distance between predictors? What family of perfectly calibrated predictors should we consider? Tackling these questions leads to a surprisingly subtle and intricate theory of distance to calibration [1, 6]; we will present formal statements later in the paper. In particular, the notion of *smooth calibration error*, introduced in the early work of [12], plays a key role in this theory. The main result of [1] is that smooth calibration error is equivalent (up to constants) to a measure they call the *lower distance to calibration*, and this is information-theoretically the best efficient approximation to the distance to calibration that one can hope to achieve.

Smooth calibration has several desirable properties as a calibration measure: it is efficient to estimate and Lipschitz continuous in the predictions. However, smooth calibration by itself does not give the type of omniprediction guarantees that one gets from perfect calibration; there are fairly simple examples of smoothly calibrated predictors p and proper losses where p incurs significantly worse expected loss than some post-processing $\kappa \circ p$. An elegant result of [10] shows that this situation can be remedied by adding some noise to the predictions before making decisions. They show that this results in an omniprediction guarantee for all proper loss functions, but where the hypothesis class is the space of all calibrated predictors for the distribution $\mathcal{D}$, rather than post-processings. Another difference is that the quality of the omniprediction guarantee decays as the distance to the calibrated predictor being used as a baseline increases. Thus, smooth calibration gives some omniprediction guarantees, provided we add random noise to smooth the predictions.

Other relaxations of perfect calibration that yield no-regret for post-processing have been studied in the literature. It is known that the expected calibration error (ECE) is an upper bound on the regret [13]. Recently, Hu and Wu [11] introduced the notion of calibration decision loss (CDL) which exactly captures this regret. But both these notions are known to be inefficient to estimate in the predictions-only access model where we are given random samples $(p(\mathbf{x}), \mathbf{y})$ for $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ [9].

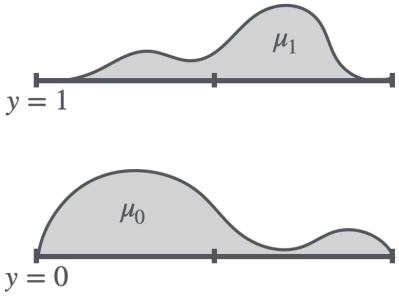


Figure 1 A typical PLD μ, with $\mu_b = \Pr_{(\mathbf{p},\mathbf{y})\sim\mu}[\mathbf{y} = b]$. Note that $\mu_0 + \mu_1 = 1$.

1.1 Preliminaries

We begin with a key new definition: *prediction-label distributions (PLDs).*

▶ **Definition 1** (Prediction-Label Distribution). *Consider the space $\mathcal{S} = [0,1] \times \{0,1\}$, where we view the first coordinate as a predicted probability $p \in [0,1]$ and the second as a label $y \in \{0,1\}$. A* Prediction-Label Distribution (PLD) *is a probability distribution μ over $\mathcal{S}$. In the case that μ has a density function, which we also denote $\mu : \mathcal{S} \to \mathbb{R}^+$, it satisfies*

$$\sum_{y\in\{0,1\}} \int_0^1 \mu(p,y)dp = 1.$$

We denote the set of all PLDs by Pld. *The space $\mathcal{S} \subseteq \mathbb{R}^2$ is equipped with the standard ℓ_1 distance. Given distributions $\mu, \nu \in$ Pld, we denote the earth mover's distance between them by $W(\mu,\nu)$. (More precisely, $W(\mu,\nu) = \inf_\pi \mathbf{E}_{(\mathbf{p},\mathbf{y},\mathbf{p}',\mathbf{y}')\sim\pi}[d((\mathbf{p},\mathbf{y}),(\mathbf{p}',\mathbf{y}'))]$, where the infimum is over all couplings π of μ with ν, and the metric is $d((p,y),(p',y')) = |p-p'| + |y-y'|$.)*

The notion of a PLD has been implicit in prior work on calibration (see for instance [1]), but the terminology is new, and reasoning directly about this space, and the earth mover's distance over it will play a key part in our results. A typical PLD is depicted in Figure 1.

▶ **Definition 2** (Calibration). *The PLD μ is (perfectly)* calibrated *if it satisfies the condition*

$$\mu(p,0)p = \mu(p,1)(1-p)$$

for all $p \in [0,1]$.[3] *We denote the set of perfectly calibrated PLDs by* Cal $\subseteq$ Pld.

Let $\mathsf{P}(\mathcal{X}) = \{p : \mathcal{X} \to [0,1]\}$ denote the space of predictors on $\mathcal{X}$. Given a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$, we associate p with a PLD $p \circ \mathcal{D}$, defined to be the distribution of the prediction-label pair $(p(\mathbf{x}), \mathbf{y})$ when $(\mathbf{x},\mathbf{y}) \sim \mathcal{D}$. We define $\mathsf{Pld}(\mathcal{D}) \subseteq \mathsf{Pld}$ to be the set of PLDs of the form $p \circ \mathcal{D}$, and $\mathsf{Cal}(\mathcal{D})$ to be the set of PLDs in $\mathsf{Pld}(\mathcal{D})$ where $p \in \mathsf{P}(\mathcal{X})$ is perfectly calibrated for the distribution $\mathcal{D}$, meaning that $\mathbf{E}[\mathbf{y}|p(\mathbf{x})] = p(\mathbf{x})$. We define $\tau(\mathcal{D}) = \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\mathbf{y} = 1]$ to be the expected label under $\mathcal{D}$. Given $\tau \in [0,1]$, we define $\mathsf{Pld}^{=\tau} \subseteq \mathsf{Pld}$ to be the subset of PLDs μ so that $\Pr_\mu[\mathbf{y} = 1] = \tau$. We further define $\mathsf{Cal}^{=\tau} = \mathsf{Pld}^{=\tau} \cap \mathsf{Cal}$ to be the subset of calibrated PLDs. We have

$$\mathsf{Cal}(\mathcal{D}) \subseteq \mathsf{Cal}^{=\tau(\mathcal{D})} \subseteq \mathsf{Cal}.$$

[3] Strictly speaking, we need this condition to hold for a set of measure 1, but we will ignore measure-theoretic subtleties throughout this paper.

Calibration error measures and loss functions are typically defined as the expectation of a suitable function on $\mathcal{S}$ under the PLD $p \circ \mathcal{D}$. For instance, writing $(\mathbf{p}, \mathbf{y}) \sim \mu$ to denote sampling according to μ, the expected calibration error (ECE) is defined as

$$\mathsf{ECE}(\mu) = \mathop{\mathbf{E}}_{(\mathbf{p},\mathbf{y})\sim\mu} \big|\mathbf{E}[\mathbf{y}|\mathbf{p}] - \mathbf{p}\big|,$$

whereas the smooth calibration error [12] is defined as

$$\mathsf{smCE}(\mu) = \max_{\psi\in\mathsf{Lip}} \mathop{\mathbf{E}}_{(\mathbf{p},\mathbf{y})\sim\mu} [\psi(\mathbf{p})(\mathbf{y} - \mathbf{p})] \tag{1}$$

where Lip is the family of 1-Lipschitz functions $\psi : [0, 1] \to [-1, 1]$. Being a property only of PLDs ensures that loss functions and calibration measures are defined uniformly for binary classification tasks across domains, regardless of whether the data are images, text, or numeric.

1.2 Prior Work

To set the stage for our work, we discuss the most relevant prior work in detail.

Distance to Calibration

Given a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$ and predictors $p, q : \mathcal{X} \to [0, 1]$, we define the expected ℓ_1 distance as

$$\ell_1^{\mathcal{D}}(p, q) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}} \big|p(\mathbf{x}) - q(\mathbf{x})\big|.$$

[1] define the *true distance to calibration* as the expected ℓ_1 distance to the closest calibrated predictor in $\mathsf{Cal}(\mathcal{D})$. However, it is hard to reason about such predictors without knowing the underlying space $\mathcal{X}$. In going from the predictor p to the PLD $p \circ \mathcal{D}$, we have lost information about the space $\mathcal{X}$ over which the predictor p is defined, which makes measuring distance to calibration challenging. The solution proposed in [1] was, given the joint distribution $\mu = p \circ \mathcal{D}$, to consider all spaces $\mathcal{X}'$, predictors $p' : \mathcal{X}' \to [0, 1]$ and distributions $\mathcal{D}'$ such that $p' \circ \mathcal{D}' = \mu$. By considering the minimum and maximum distance to calibration over all such spaces, they define two new calibration measures, the lower and upper distance to calibration, which sandwich the true distance to calibration between them. Formally:

▶ **Definition 3.** *Given a PLD μ, let* $\mathrm{lift}(\mu)$ *be the set of all predictor-distribution pairs $(p', \mathcal{D}')$ such that $p' \circ \mathcal{D}' = \mu$.*[4] *Given a predictor $p : \mathcal{X} \to [0, 1]$ and a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$,*

- *The (true) distance to calibration is defined as*

$$\mathsf{dCE}_{\mathcal{D}}(p) = \inf_{\substack{q:\mathcal{X}\to[0,1] \\ q\circ\mathcal{D}\in\mathsf{Cal}(\mathcal{D})}} \ell_1^{\mathcal{D}}(p, q).$$

- *The upper distance to calibration is defined as*

$$\overline{\mathsf{dCE}}(p \circ \mathcal{D}) = \sup_{(p',\mathcal{D}')\in\mathrm{lift}(p\circ\mathcal{D})} \mathsf{dCE}_{\mathcal{D}'}(p').$$

- *The lower distance to calibration is defined as*

$$\underline{\mathsf{dCE}}(p \circ \mathcal{D}) = \inf_{(p',\mathcal{D}')\in\mathrm{lift}(p\circ\mathcal{D})} \mathsf{dCE}_{\mathcal{D}'}(p').$$

[4] The underlying *feature space* $\mathcal{X}'$ over which p' and $\mathcal{D}'$ are defined can vary for the elements in $\mathrm{lift}(\mu)$.

Observe that unlike dCE, both $\overline{\mathsf{dCE}}$ and $\underline{\mathsf{dCE}}$ are only functions of the PLD $p \circ \mathcal{D}$, and not the underlying domain $\mathcal{X}$. From the definitions, it follows that

$$\underline{\mathsf{dCE}}(p \circ \mathcal{D}) \leq \mathsf{dCE}_{\mathcal{D}}(p) \leq \overline{\mathsf{dCE}}(p \circ \mathcal{D}).$$

[1] ask how easy it is to compute each of the quantities $\underline{\mathsf{dCE}}$, dCE, and $\overline{\mathsf{dCE}}$ from sample access to $p \circ \mathcal{D}$. They show that the notion of smooth calibration error from [12] (defined in Equation (1)) holds the key to the answer.

▶ **Theorem 4** ([1]). *$1/2 \leq \underline{\mathsf{dCE}}(\mu)/\mathsf{smCE}(\mu) \leq 2$ for every PLD μ.*

Further, [1] show that the upper and lower distance are within a quadratic factor of each other:

$$\overline{\mathsf{dCE}}(\mu) \leq 4\sqrt{\underline{\mathsf{dCE}}(\mu)}, \tag{2}$$

and that it is information theoretically impossible to estimate the true distance better than within a quadratic factor. This tells us that $\mathsf{smCE}(p \circ \mathcal{D})$ is essentially the best approximation one can get to both the true distance $\mathsf{dCE}_{\mathcal{D}}(p)$, and the lower distance $\underline{\mathsf{dCE}}(p \circ \mathcal{D})$ (up to constant factors). They leave open the question of whether we can compute the upper distance $\overline{\mathsf{dCE}}(p \circ \mathcal{D})$ any better.

Omniprediction From Smooth Calibration

A (bounded) loss function is a function $\ell : \mathcal{S} \to [-1, 1]$, and a proper loss is one where if the label $\mathbf{y} \sim \mathrm{Ber}(p)$, then $\mathbf{E}[\ell(q, \mathbf{y})]$ is minimized at $q = p$. We let $\mathcal{L}$ denote the family of all bounded proper losses, and define the notion of omnipredictors, specialized to this family.

▶ **Definition 5** (Omnipredictor, [8]). *Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{0, 1\}$, let $\mathcal{L}$ be the family of bounded proper losses, and $\mathcal{Q} \subseteq \mathsf{P}(\mathcal{X})$ be a set of predictors. A predictor $p : \mathcal{X} \to [0, 1]$ is an $(\mathcal{L}, \mathcal{Q}, \alpha)$-omnipredictor if for every $\ell \in \mathcal{L}$ and $q \in \mathcal{Q}$, it holds that*

$$\mathbf{E}[\ell(p(\mathbf{x}), \mathbf{y})] \leq \mathbf{E}[\ell(q(\mathbf{x}), \mathbf{y})] + \alpha.$$

The setting $\alpha = 0$ corresponds to perfect omniprediction, where p is optimal compared to $\mathcal{Q}$, while $\alpha = 2$ is trivial since our losses are bounded in $[-1, 1]$. So we will consider $\alpha \in [0, 2)$. We are interested in settings where one can have α bounded away from 2, ideally tending to 0.[5]

Foster and Vohra [5] showed that a perfectly calibrated predictor p is a perfect omnipredictor for all proper losses.

▶ **Theorem 6** ([5]). *Let p be perfectly calibrated. Let $K = \{\kappa \mid \kappa : [0, 1] \to [0, 1]\}$ be all possible post-processings. Then for any $\ell \in \mathcal{L}$, and $\kappa \in K$ we have $\mathbf{E}[\ell(\mathbf{p}, \mathbf{y})] \leq \mathbf{E}[\ell(\kappa(\mathbf{p}), \mathbf{y})]$.*[6]

One cannot replace perfect calibration with smooth calibration in their statement: For every $\varepsilon > 0$, we will give an example of a PLD μ with $\mathsf{smCE}(\mu) \leq \varepsilon$, a loss $\ell \in \mathcal{L}$, and post-processing κ so that

$$\mathop{\mathbf{E}}_{\mu}[\ell(\mathbf{p}, \mathbf{y})] = 1, \quad \mathop{\mathbf{E}}_{\mu}[\ell(\kappa(\mathbf{p}), \mathbf{y})] = 0,$$

[5] Like calibration, expected loss is a property of a PLD, rather than the predictor itself. However, in omniprediction, the underlying distribution $\mathcal{D}$ is fixed, which fixes the marginal distribution on $\mathbf{y}$. Hence, we will think of omniprediction both as a property of PLDs that lie in $\mathsf{Pld}^{=\tau(\mathcal{D})}$, and of predictors $p \in \mathsf{P}(\mathcal{X})$, with the associated PLD $p \circ \mathcal{D}$.

[6] They prove this for the squared loss, but it extends straightforwardly to all proper losses.

This raises the question of whether it is at all possible to show omniprediction guarantees against a rich class of post-processings for a smoothly calibrated predictor.

A novel omniprediction guarantee for smooth calibration error was shown in the recent work of [10]. There are two important ways in which their guarantee differs from the Foster-Vohra result:

1. **Smoothed predictions.** Rather than using the predictions $\mathbf{p}$ of a smoothly calibrated predictor directly, they add some noise $\mathbf{z}$ sampled uniformly from the interval $[-\sigma, \sigma]$ and truncate the result to $[0, 1]$. We will denote this noise distribution by $\mathbf{z} \sim [\pm\sigma]$ and the truncated predictor by $\mathbf{p}_{\mathbf{z}}$.
2. **The baseline class.** The omniprediction guarantee holds for the baseline class of all calibrated PLDs $\mu \in \mathsf{Cal}^{=\tau(\mathcal{D})}$, which satisfy accuracy in expectation. This is a rich class that includes the Bayes optimal PLD $\mu^* = (\mathbf{E}[\mathbf{y}|\mathbf{x}], \mathbf{y})$, and it is impossible to hope for an omniprediction guarantee with any fixed $\alpha < 1$. Instead, they allow α to degrade with the earth mover's distance between $p \circ \mathcal{D}$ and μ.

Formally, [10] show that smoothing p gives an omnipredictor for $\mathsf{Cal}^{=\tau(\mathcal{D})}$ with the following guarantee:[7]

▶ **Theorem 7** ([10]). *Let $p \in \mathsf{P}(\mathcal{X})$, $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{0, 1\}$ and $\nu \in \mathsf{Cal}^{=\tau(\mathcal{D})}$. For $\ell \in \mathcal{L}$ and $\sigma \in (0, 1]$, we have*

$$\mathop{\mathbf{E}}_{\substack{(\mathbf{p},\mathbf{y})\sim p\circ\mathcal{D},\\ \mathbf{z}\sim[\pm\sigma]}}[\ell(\mathbf{p}_{\mathbf{z}}, \mathbf{y})] \leq \mathop{\mathbf{E}}_{(\mathbf{q},\mathbf{y})\sim\nu}[\ell(\mathbf{q}, \mathbf{y})] + O\left(\sigma + \frac{1}{\sigma} W(\mathbf{p} \circ \mathcal{D}, \nu)\right).$$

To connect this to smooth calibration, we observe the smallest that the earth mover's distance can be is $\underline{\mathsf{dCE}}(p) = \Theta(\mathsf{smCE}(p))$, by the result of [1]. We can take $\sigma = \sqrt{\mathsf{smCE}(p \circ \mathcal{D})}$, to get an omniprediction error bound of $\alpha = O(\sqrt{\mathsf{smCE}(p \circ \mathcal{D})})$. Thus the guarantee is particularly meaningful for smoothly calibrated predictors, where α goes to 0. Since smooth calibration error can be efficiently estimated, one can estimate the level of noise to add to the predictor before making decisions.

Given these two incomparable results, there are two natural questions that arise:

- Can one extend the Foster-Vohra guarantee (Theorem 6) and show an omniprediction guarantee for smoothly calibrated predictors, against the class of post-processings?
- Can one relax the assumption of calibration on $\mathbf{q}$ in Theorem 7 and show an omniprediction guarantee against all PLDs in $\mathsf{Pld}^{=\tau(\mathcal{D})}$?

1.3 Our Results

In this work, we prove some new results and reprove some old results about the properties of smooth calibration, both as a replacement for perfect calibration that gives strong omniprediction guarantees, and as a measure of the distance to calibration.

- We present a general omniprediction result for smooth calibration, that strengthens and generalizes the Foster-Vohra result by allowing post-processings and the [10] result by comparing with all nearby predictors, whether or not they are calibrated.

[7] The exact statement of Theorem 7 does not appear in [10], but it can be derived from Theorem 3.1 therein. There are some differences between their statement and ours. They sample the noise from a suitably chosen DP mechanism (Gaussian or Laplace), but we prefer random shifts because they are simpler, and as we will see, they correspond to random bucketing which is commonly used in practice. Although their bound is not stated in terms of $W(\cdot, \cdot)$, the way we have stated it is equivalent by Lemma 15.

- We present a crisp new characterization of the lower distance to calibration in terms of earthmover distance from the set Cal. We use this to present an arguably simpler proof of the tight connection between smooth calibration error and the lower distance to calibration.
- We show that the upper distance to calibration cannot be estimated to better than a quadratic factor, with sample complexity independent of the support size of the predictor.

Common to all our results is a view of calibrated predictors on a space being sandwiched between the larger set of all calibrated distributions on prediction-label pairs, and the smaller set of calibrated post-processings, which is loosely inspired by convex relaxations in discrete optimization. This view informs our formulation of theorem statements, we believe it also results in more general and clarifying proofs.

2 Technical Overview

In this section, we present full technical statements of our results and the intuition behind them, without getting into proof details.

2.1 Omniprediction From Smooth Calibration

We show an omniprediction guarantee for smoothly calibrated predictors, against the baseline class of post-processings of predictors in $\mathsf{PId}^{=\tau(\mathcal{D})}$.[8] We write $f \lesssim g$ if $f = O(g)$ and p_z for $[p+z]_0^1$, where $[\cdot]_0^1$ denotes projection onto the interval $[0,1]$.

▶ **Theorem 8.** *Let $p \in \mathsf{P}(\mathcal{X})$, $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{0,1\}$, $\mu = p \circ \mathcal{D}$ and $\nu \in \mathsf{PId}^{=\tau(\mathcal{D})}$. For any bounded proper loss $\ell \in \mathcal{L}$, post-processing function $\kappa : [0,1] \to [0,1]$, and $\sigma \in (0,1]$,*

$$\mathop{\mathbf{E}}_{\substack{(\mathbf{p},\mathbf{y})\sim\mu \\ \mathbf{z}\sim[\pm\sigma]}}[\ell(\mathbf{p_z},\mathbf{y})] - \mathop{\mathbf{E}}_{\substack{(\mathbf{q},\mathbf{y})\sim\nu \\ \mathbf{z}\sim[\pm\sigma]}}[\ell(\kappa(\mathbf{q_z}),\mathbf{y})] \lesssim \sigma + \frac{1}{\sigma}\Big(\mathsf{smCE}(\mu) + W(\mu,\nu)\Big).$$

Our theorem allows for arbitrary PLDs $\nu \in \mathsf{PId}^{=\tau(\mathcal{D})}$, allowing the guarantee to decay with $W(\mu,\nu)$. It also allows for arbitrary post-processings κ. Thus it gives a common generalization of the baseline classes used in [5] and [10].

In the setting where $\mathbf{q} = \mathbf{p}$, our result implies the following bound:

▶ **Corollary 9.** *Let $(\mathbf{p},\mathbf{y}) \sim \mu$ and let $\mathbf{z}$ be uniform over $[-\sigma,\sigma]$ and independent from $(\mathbf{p},\mathbf{y})$. Then, the following is true for any proper loss $\ell : [0,1] \times \{0,1\} \to [-1,1]$ and any $\kappa : [0,1] \to [0,1]$:*

$$\mathop{\mathbf{E}}_{\substack{(\mathbf{p},\mathbf{y})\sim\mu \\ \mathbf{z}\sim[\pm\sigma]}}[\ell(\mathbf{p_z},\mathbf{y}) - \ell(\kappa(\mathbf{p_z}),\mathbf{y})] \lesssim \sigma + \frac{\mathsf{smCE}(\mu)}{\sigma}. \tag{3}$$

This can be interpreted as saying that smoothly calibrated predictors are omnipredictors with respect to arbitrary post-processings, provided we add noise to the predictions. This gives a *smoothed analysis* analogue for the Foster-Vohra result for smooth calibration. It is essential to smooth the comparison baseline $\mathbf{q}$ with random noise, the above statement is not true if we replace $\kappa(\mathbf{p_z})$ with $\kappa(\mathbf{p})$, and only assume that p is smoothly calibrated.

[8] The guarantee stated would also hold for $\nu \in \mathsf{PId}$, but changing the label distribution is not natural for omniprediction.

A similar statement may be derived by combining the results of Blasiok and Nakkiran [3], showing that adding noise (from a different distribution) to a smoothly calibrated predictor yields small ECE, with the results of [13, 11], showing that small ECE suffices for omniprediction guarantees against post-processings.

In the setting of [10], where we restrict the benchmark class to calibrated PLDs from $\mathsf{Cal}^{=\tau(\mathcal{D})}$, we do not need to smooth $\mathbf{q}$, and can show a quantitatively stronger bound:

▶ **Theorem 10.** *Let $p \in \mathsf{P}(\mathcal{X})$, $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{0,1\}$, $\mu = p \circ \mathcal{D}$ and $\nu \in \mathsf{Cal}^{=\tau(\mathcal{D})}$. For any bounded proper loss $\ell \in \mathcal{L}$, and $\sigma \in (0,1]$*

$$\mathop{\mathbf{E}}_{\substack{(\mathbf{p},\mathbf{y})\sim p\circ\mathcal{D}\\ \mathbf{z}\sim[\pm\sigma]}}[\ell(\mathbf{p}_{\mathbf{z}},\mathbf{y})] - \mathop{\mathbf{E}}_{(\mathbf{q},\mathbf{y})\sim\nu}[\ell(\mathbf{q},\mathbf{y})] \lesssim \sigma + \frac{\mathsf{smCE}(\mu)}{\sigma} + W(\mu,\nu).$$

Since ν is calibrated, post-processing does not reduce the loss. Hence the same bound holds for all post-processings $\kappa(\mathbf{q})$. In comparison to Theorem 7, our result replaces the term $W(\mu,\nu)/\sigma$ with $\mathsf{smCE}(\mu)/\sigma + W(\mu,\nu)$. Corollary 16 shows that $\mathsf{smCE}(\mu)$ is within constant factors of the minimum value of $W(\mu,\nu)$ over all $\nu \in \mathsf{Cal}^{=\tau(\mathcal{D})}$. For such ν, the two bounds are similar up to constants. But when $W(\mu,\nu) \gg \mathsf{smCE}(\mu)$, we incur the smaller penalty of $W(\mu,\nu)$ rather than $W(\mu,\nu)/\sigma$, thus improving over Theorem 7.

Our proof uses techniques from the literature on loss outcome-indistinguishability, introduced in [7] and used in the calibration context by [9]. However, we believe a key contribution is identifying the right statement itself, which is rather delicate. To illustrate this, consider a simple example: Consider a two point space $\mathcal{X} = \{x_0, x_1\}$. The distribution $\mathcal{D}$ is uniform on $(x_0, 0)$ and $(x_1, 1)$. We will consider the *proper* version of 0-1 loss, $\ell_{\{0,1\}} : [0,1] \times \{0,1\} \to \{0,1\}$ where

$$\ell_{\{0,1\}}(p,y) = |1(p \geq 1/2) - y|.$$

In other words, if we are on the correct side of $1/2$, we suffer 0 loss, else the loss is 1. This is essentially a special case of the v-shaped losses which are studied in the literature. Now consider the following predictors:

- The *good* predictor g where $g(x_0) = 1/2 - \varepsilon, g(x_1) = 1/2 + \varepsilon$, which has expected loss 0.
- The *bad* predictor b where $b(x_0) = 1/2 + \varepsilon, b(x_1) = 1/2 - \varepsilon$, which has expected loss 1.
- The *uniform* predictor u where $u(x_0) = u(x_1) = 1/2$, which has expected loss $1/2$.

Smooth calibration error does not distinguish between good and bad (or even uniform), since $\mathsf{smCE}(b) = \mathsf{smCE}(g) = \varepsilon$, while u is perfectly calibrated. We observe that for $\mathbf{z} \in [\pm\sigma]$, the total variation distance between $g_{\mathbf{z}}$ and $b_{\mathbf{z}}$ is $\mathrm{dTV}(g_{\mathbf{z}}, b_{\mathbf{z}}) \leq 2\varepsilon/\sigma$. Finally, $g = \kappa(b)$ where $\kappa(p) = 1 - p$. With these observations in hand, we can deduce the following (see the full version for more details):

- The gap between expected loss of p and $\kappa(p)$ can be 1, even when $\mathsf{smCE}(p) \leq \varepsilon$. To see this, take $p = b$, $q = \kappa(p) = g$. Thus one cannot hope for a direct analogue of the Foster-Vohra bound for smooth calibration without any smoothing.
- The same example shows that the gap between the expected loss of $p_{\mathbf{z}}$ and $q = \kappa(p)$ can be $1/2$, even when $\mathsf{smCE}(p) \leq \varepsilon$, and $W(p \circ \mathcal{D}, q \circ \mathcal{D}) \leq \varepsilon$. Note that when $\sigma \gg \varepsilon$, $p_{\mathbf{z}}$ is essentially the same as u, so it has expected loss $1/2$. Thus smoothing p alone but not smoothing q is not sufficient.
- By taking $p = b$ and $q = g$, we see that the gap between the expected loss of p and $q_{\mathbf{z}}$ can be $1/2$, since now $q_{\mathbf{z}}$ is essentially the same as u for $\sigma \gg \varepsilon$. So smoothing q alone without smoothing p is also not sufficient.

Further, in the full version, we give examples which illustrate that the omniprediction error must scale linearly with each of the three terms on the RHS of Theorem 8.

2.2 Lower Distance to Calibration

Our next set of results uses what we call the *earth mover's distance to calibration*, denoted dEMC, to prove new results regarding the upper and lower distance to calibration, as well as provide new proofs of existing results. We begin with the definition of dEMC:

▶ **Definition 11** (dEMC). *Given $\mu \in \mathsf{Pld}$, the* earth mover's distance to calibration *is*

$$\mathsf{dEMC}(\mu) = \inf_{\nu \in \mathsf{Cal}} W(\mu, \nu).$$

To begin, we use dEMC to give a simple proof of Theorem 4, which states that the smooth calibration error approximates the lower distance to calibration up to a constant factor. This result originally appeared in [1], with an alternate proof in [3]. Next, we show that the upper distance to calibration is hard to approximate within a quadratic factor in the *prediction-only access model*, where we have access to samples of the form $(\mathbf{p}, \mathbf{y})$. Specifically, Theorem 20 is a sample complexity lower bound that scales with $\Omega(\sqrt{k})$, where k is the support size of the prediction distribution. This result complements prior work showing that the (ordinary) distance to calibration is *impossible* to estimate from samples within a quadratic factor in the same model [1].

Lower Distance and Earth Mover's Distance

Our new proof of Theorem 4 is based on the following two lemmas. The first lemma relates smooth calibration error to the earth mover's distance to calibration, which the second lemma in turn equates to the lower distance to calibration. Interestingly, the second lemma can be viewed as an alternate and natural *definition* of the lower distance to calibration.

▶ **Lemma 12.** *$1/2 \le \mathsf{dEMC}(\mu)/\mathsf{smCE}(\mu) \le 2$ for every PLD μ.*

▶ **Lemma 13.** *$\underline{\mathsf{dCE}}(\mu) = \mathsf{dEMC}(\mu)$ for every PLD μ.*

It is clear that these two results yield Theorem 4 when combined. It remains to prove them. The crucial step in the proof of Lemma 12 uses the apparent flexibility in the definition of dEMC, which allows one to change the $\mathbf{y}$ distribution when designing a nearby calibrated predictor. In particular, given $\mu = (\mathbf{p}, \mathbf{y})$, it is easy to see that $\nu = (\mathbf{p}, \tilde{\mathbf{y}})$ where $\tilde{\mathbf{y}} \sim \mathrm{Ber}(\mathbf{p})$ is calibrated, and $W(\mu, \nu) = \Theta(\mathsf{smCE}(\mu))$ (see [6, Lemma 3.4]). The proof of Lemma 13, in contrast, shows that this flexibility in the $\mathbf{y}$ distribution, while convenient, was not actually necessary and that we can restrict ν to lie in the set $\mathsf{Cal}^{=\tau}$ where $\tau = \Pr_\mu[\mathbf{y} = 1]$.

In contrast to our proof strategy, which centers on the earth mover's distance to the set Cal, it turns out the prior proof of [1] was effectively considering $\mathsf{Cal}^{=\tau}$. To see this connection, we use a helpful interpretation of $\underline{\mathsf{dCE}}$, proposed in [1], as the following infimum:

▶ **Lemma 14** ([1]). *$\underline{\mathsf{dCE}}(\mu) = \inf \mathbf{E}|\mathbf{p} - \mathbf{q}|$, where the infimum is taken over all triples of joint random variables $(\mathbf{p}, \mathbf{q}, \mathbf{y})$ such that $(\mathbf{p}, \mathbf{y}) \sim \mu$ and $(\mathbf{q}, \mathbf{y})$ is calibrated.*

The next lemma (proved in Section 4) shows how to move from such triples to earth mover's distance.

▶ **Lemma 15.** *For any two $\mu, \nu \in \mathsf{Pld}^{=\tau}$ for $\tau \in [0, 1]$, we have that $W(\mu, \nu) = \inf \mathbf{E}[|\mathbf{p} - \mathbf{q}|]$, where the infimum is taken over all triples of joint random variables $(\mathbf{p}, \mathbf{q}, \mathbf{y})$ such that $(\mathbf{p}, \mathbf{y}) \sim \mu$ and $(\mathbf{q}, \mathbf{y}) \sim \nu$.*

Putting these two lemmas together, it follows that the prior characterization of $\underline{\mathsf{dCE}}$ can be viewed as the earth mover's distance to $\mathsf{Cal}^{=\tau}$ for an appropriate choice of τ:

▶ **Corollary 16.** $\underline{\mathsf{dCE}}(\mu) = \inf_{\nu \in \mathsf{Cal}^{=\tau}} W(\mu, \nu)$, *where* $\tau = \Pr_{(\mathbf{p},\mathbf{y})\sim\mu}[\mathbf{y} = 1]$.

As the proof in the present paper shows, working with the earth mover's distance to Cal *directly*, rather than to $\mathsf{Cal}^{=\tau}$, is equivalent and leads to a cleaner picture. Furthermore, the original proof of the lower bound of Theorem 4 from [1] used an argument based on linear programming duality, which in hindsight resembles a modified version Kantorovich-Rubinstein duality for earth mover's distance. Our proof seems to circumvent this argument by directly using Kantorovich-Rubinstein duality, which we of course need not rederive, in the proof of Lemma 12. The results of this section also yield the following interesting corollary, which we state purely in terms of earth mover's distance to calibration.

▶ **Corollary 17.** *Given a PLD μ, the following are equal up to constant factors:*

(a) *Its earth mover's distance to calibration,* $\mathsf{dEMC}(\mu)$*:*

$$\inf\left\{W(\mu,\nu) \,\middle|\, \nu \in \mathsf{Cal}\right\},$$

(b) *Its earth mover's distance to calibration while preserving the marginal of* $\mathbf{y}$*:*

$$\inf\left\{W(\mu,\nu) \,\middle|\, \nu \in \mathsf{Cal}^{=\tau}\right\},$$

where $\tau = \Pr_{(\mathbf{p},\mathbf{y})\sim\mu}[\mathbf{y} = 1]$.

(c) *Its earth mover's distance to calibration while preserving the marginal of* $\mathbf{p}$*:*

$$\inf\left\{W(\mu,\nu) \,\middle|\, \nu \in \mathsf{Cal} \text{ and } \Pr_{(\mathbf{p},\mathbf{y})\sim\mu}[\mathbf{p} \le t] = \Pr_{(\mathbf{p},\mathbf{y})\sim\nu}[\mathbf{p} \le t] \text{ for all } t \in [0,1]\right\}.$$

More precisely, $a = b \le c \le 2a$.

2.3 Intractability of the Upper Distance

We have already defined the sets Pld and $\mathsf{Pld}^{=\tau(\mathcal{D})}$, which are supersets of the set $\mathsf{Pld}(\mathcal{D})$ of PLDs induced by $\mathsf{P}(\mathcal{X})$. We now define the set of post-processings of a predictor, which are a subset of the space $\mathsf{P}(\mathcal{X})$ that will be relevant for our discussion of upper distance to calibration.

▶ **Definition 18** (Post-processings of predictors and PLDs)**.** *Let* $K = \{\kappa : [0,1] \to [0,1]\}$ *denote the set of all possible post-processing functions. Let* $K(p) = \{\kappa \circ p\}_{\kappa\in K}$ *denote all post-processings of a predictor* p*, and observe that* $K(p) \subseteq \mathsf{P}(\mathcal{X})$*. Given a PLD* $\mu = (\mathbf{p}, \mathbf{y})$*, let* $K(\mu)$ *denote the set of all PLDs of the form* $(\kappa(\mathbf{p}), \mathbf{y})_{\kappa\in K}$*, and let* $\mathsf{Cal}(\mu) = \mathsf{Cal} \cap K(\mu)$.

We have the following inclusions among the set of PLDs for any predictor p and distribution $\mathcal{D}$:

$$\mathsf{Pld}(\mathcal{D}) \supseteq K(p \circ \mathcal{D}),\ \mathsf{Cal}(\mathcal{D}) \supseteq \mathsf{Cal}(p \circ \mathcal{D}).$$

In order to prove the existence of a PLD in $\mathsf{Pld}(\mathcal{D})$ with a certain property, one typically proves the existence of a PLD in $K(p \circ \mathcal{D})$. This is equivalent to only considering predictors in $K(p)$ rather than all predictors from $\mathsf{P}(\mathcal{X})$. However, K is not a particularly tractable set, since it consists of all post-processing functions. Optimizing over it efficiently can be hard, as was recently shown in the context of Calibration Decision Loss by [9]. We show a similar barrier to efficiently estimating the upper distance to calibration.

We show that accurate estimates of the upper distance to calibration, which are asymptotically better than what is implied by the quadratic relation to the lower distance (Equation (2)) are hard to obtain in the prediction-only access model. Specifically, we prove a sample

complexity lower bound that scales with the size of the support of the distribution, which may be very large, even if finite. Our result complements prior work showing that the (ordinary) distance to calibration is not just hard, but *impossible* to estimate in this model with the same level of accuracy. As we shall soon see, both of these results are most easily understood from the perspective of the earth mover's distance to calibration.

A key concept in the proof of the result of this section, as well as later results in this paper, is the following *almost balanced* PLD. The distribution corresponds to a predictor which always predicts $1/2 \pm \varepsilon$. The direction of the deviation from $1/2$, however, is uncorrelated with the true label, which is a uniformly random bit.

▶ **Definition 19** (Almost Balanced PLD). *Given $\varepsilon > 0$, the* ε-almost balanced PLD *is the distribution on $\mathcal{S} = [0,1] \times \{0,1\}$ with mass $1/4$ on each of $(1/2-\varepsilon, 0)$, $(1/2-\varepsilon, 1)$, $(1/2+\varepsilon, 0)$, and $(1/2+\varepsilon, 1)$.*

Our next result will construct several examples, each comprising a domain $\mathcal{X}$, a predictor $p : \mathcal{X} \to [0,1]$, and a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$. Roughly speaking, in each example, the prediction-label distribution $p \circ \mathcal{D}$ is similar to the ε-almost balanced PLD, but distinguishing the scenarios from samples of the form $(p(\mathbf{x}), \mathbf{y})$ may be hard. Our proof of the theorem will make use of intuitive visual arguments enabled by the earth mover's perspective. Before we state the theorem, we briefly recall the following view of $\overline{\mathsf{dCE}}$ in terms of $\mathsf{Cal}(\mu)$ from [1]:

$$\overline{\mathsf{dCE}}(\mu) = \inf_{\kappa} \mathop{\mathbf{E}}_{(\mathbf{p},\mathbf{y})\sim\mu} |\kappa(\mathbf{p}) - \mathbf{p}|,$$

where $\kappa : [0,1] \to [0,1]$ ranges over all calibrated post-processings, meaning $\mathbf{E}[\mathbf{y}|\kappa(\mathbf{p})] = \kappa(\mathbf{p})$.

▶ **Theorem 20** (Succinct Version of Theorem 26). *Fix parameters $\varepsilon > 0$, $k \in \mathbb{N}$, and $\delta_{ij} \in \mathbb{R}$ for $(i,j) \in \{0,1\} \times [k]$. There exist four 4-tuples $(\mathcal{X}_a, \mathcal{D}_{\mathcal{X}_a}, p^*_a, p_a)$, $(\mathcal{X}_b, \mathcal{D}_{\mathcal{X}_b}, p^*_b, p_b)$, $(\mathcal{X}_c, \mathcal{D}_{\mathcal{X}_c}, p^*_c, p_c)$, and $(\mathcal{X}_d, \mathcal{D}_{\mathcal{X}_d}, p^*_d, p_d)$ with the following properties. First, $|\mathcal{X}_c| = |\mathcal{X}_d| = k$. Second, cases (a) and (b) do not depend on the δ_{ij} parameters. Finally, if δ_{ij} are sampled i.i.d. from an appropriate continuous distribution, then:*

- **(i)** *Case (a), which has $\mathsf{dCE} \geq \Omega(\varepsilon)$, is impossible to distinguish with any positive advantage from case (b), which has $\mathsf{dCE} \leq O(\varepsilon^2)$, given any finite number of prediction-label samples.*
- **(ii)** *It requires at least $\Omega(\sqrt{k})$ prediction-label samples to distinguish (with constant advantage in expectation over the choice of δ_{ij}) case (c), which has $\overline{\mathsf{dCE}} \geq \Omega(\varepsilon)$, from case (d), which has $\overline{\mathsf{dCE}} \leq O(\varepsilon^2)$ (with probability 1 over the choice of δ_{ij}).*

In particular, dCE cannot be estimated within a better-than-quadratic factor from prediction-label samples, and $\overline{\mathsf{dCE}}$ cannot be estimated within a better-than-quadratic factor from any number of prediction-label samples that is independent of the support size of the distribution of predictions.

Observe that part (i) of Theorem 26 is precisely the prior result of [1] regarding the inapproximability of dCE in the prediction-only access model. Part (ii) is our new result regarding $\overline{\mathsf{dCE}}$. Working with PLDs and earth mover's distance greatly simplifies the task of finding explicit examples which exhibit separations between various measures; rather than conjuring up a clever space $\mathcal{X}'$, we simply find PLDs in Pld or $K(p \circ \mathcal{D})$ that exhibit the desired separation.

3 Omniprediction From Smooth Calibration

In this section, we prove Theorems 8. Recall that $\mathbf{z} \sim [\pm\sigma]$ denotes $\mathbf{z} \sim \mathrm{Unif}[-\sigma, \sigma]$. Both theorems consider the smoothed predictor $p_{\mathbf{z}} = [p + \mathbf{z}]_0^1$, An equivalent way to interpret the smoothing operation is as follows. Draw $\mathbf{w} \sim \mathrm{Unif}[0, 2\sigma]$ and round p to the nearest point in the randomly shifted grid

$$\mathcal{I}_{\mathbf{w}} = \{0, 1\} \cup \{[\mathbf{w} + 2i\sigma]_0^1 : i \in \mathbb{Z}\}.$$

For any fixed $p \in [0, 1]$, the two procedures produce random variables with the same distribution.

We start with some simple lemmas.

▶ **Lemma 21.** *Let $z \in [-\sigma, \sigma]$ and $p \in [0, 1]$. Then we have $d_{\mathrm{TV}}(\mathrm{Bern}(p), \mathrm{Bern}(p_z)) \le \sigma$.*

We skip the simple proof. The next lemma shows that adding noise makes bounded functions Lipschitz.

▶ **Lemma 22.** *If $f : [0, 1] \to [-1, 1]$ is measurable and $\mathbf{z}$ is uniform in $[-\sigma, \sigma]$, then the function f_σ with $f_\sigma(p) = \mathbf{E}[f(p_{\mathbf{z}})]$ is $O(1/\sigma)$-Lipschitz in p.*

Proof. Let $p, q \in [0, 1]$. If $\mathbf{z} \sim [-\sigma, \sigma]$, then the total variation distance between the random variables $p_{\mathbf{z}}$ and $q_{\mathbf{z}}$ is $O(|p - q|/\sigma)$. This is because the density of $p + \mathbf{z}$ is $D_1(t) = \mathbf{1}[|t - p| \le \sigma]/2\sigma$ and the density of $q + \mathbf{z}$ is $D_2(t) = \mathbf{1}[|t - q| \le \sigma]/2\sigma$. Therefore:

$$d_{\mathrm{TV}}(p + \mathbf{z}, q + \mathbf{z}) = \frac{1}{2} \int_{t=-\infty}^{\infty} |D_1(t) - D_2(t)|\, dt \lesssim \frac{|p - q|}{\sigma}.$$

The result follows from the fact that post-processing (i.e., clipping and applying f) does not increase the total variation distance. ◀

We now prove our main omniprediction result for smooth calibration, which we restate below.

▶ **Theorem 8.** *Let $p \in \mathsf{P}(\mathcal{X})$, $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{0, 1\}$, $\mu = p \circ \mathcal{D}$ and $\nu \in \mathsf{PId}^{=\tau(\mathcal{D})}$. For any bounded proper loss $\ell \in \mathcal{L}$, post-processing function $\kappa : [0, 1] \to [0, 1]$, and $\sigma \in (0, 1]$,*

$$\mathop{\mathbf{E}}_{\substack{(\mathbf{p},\mathbf{y})\sim\mu \\ \mathbf{z}\sim[\pm\sigma]}} [\ell(\mathbf{p}_{\mathbf{z}}, \mathbf{y})] - \mathop{\mathbf{E}}_{\substack{(\mathbf{q},\mathbf{y})\sim\nu \\ \mathbf{z}\sim[\pm\sigma]}} [\ell(\kappa(\mathbf{q}_{\mathbf{z}}), \mathbf{y})] \lesssim \sigma + \frac{1}{\sigma}\Big(\mathsf{smCE}(\mu) + W(\mu, \nu)\Big).$$

We use the notation $\mathbf{y}^p$ to denote the random variable following the Bernoulli distribution with probability of success p. Our proof will follow the loss OI technique introduced by [7], where we start from a label distribution where the desired omniprediction guarantee holds true by Bayes optimality, and then bound the cost of modifying the label distribution. Since the predictor we wish to show omniprediction guarantees for is $\mathbf{p}_{\mathbf{z}}$, the appropriate label distribution is $\mathbf{y}^{\mathbf{p}_{\mathbf{z}}}$ where $\mathbf{E}[\mathbf{y}^{\mathbf{p}_{\mathbf{z}}} | \mathbf{p}_{\mathbf{z}}] = \mathbf{p}_{\mathbf{z}}$. Switching between $\mathbf{y}^{\mathbf{p}_{\mathbf{z}}}$ and $\mathbf{y}$ is enabled by the following lemma:

▶ **Lemma 23.** *Let $\ell : [0, 1] \times \{0, 1\} \to [-1, 1]$ be a bounded, but not necessarily proper loss function. Then*

$$|\mathbf{E}[\ell(\mathbf{p}_{\mathbf{z}}, \mathbf{y})] - \mathbf{E}[\ell(\mathbf{p}_{\mathbf{z}}, \mathbf{y}^{\mathbf{p}_{\mathbf{z}}})]| \lesssim \sigma + \frac{\mathsf{smCE}(\mu)}{\sigma}. \tag{4}$$

Proof. Following [7], defining $\partial\ell(p) = \ell(p,1) - \ell(p,0)$, we can write

$$\ell(p,y) = \ell(p,0) + y\partial\ell(p).$$

Define the function $w(p) = \mathbf{E}_{\mathbf{z}}[\partial\ell(p_{\mathbf{z}})]$. Since $|\ell| \le 1$, $|\partial\ell| \le 2$. Hence Lemma 22 implies that $w(p)$ is $O(1/\sigma)$-Lipschitz. Hence we have

$$\left|\mathbf{E}[\ell(\mathbf{p_z}, \mathbf{y})] - \mathbf{E}[\ell(\mathbf{p_z}, \mathbf{y^p})]\right| = \mathbf{E}[\partial\ell(\mathbf{p_z})(\mathbf{y} - \mathbf{y}^p)] = \mathbf{E}[w(\mathbf{p})(\mathbf{y} - \mathbf{p})] \le O\left(\frac{\mathsf{smCE}(\mu)}{\sigma}\right). \tag{5}$$

where the last inequality is from the definition of smooth calibration, since $\sigma w(p)$ is 1-Lipschitz. Further we have

$$|\mathbf{E}[\ell(\mathbf{p_z}, \mathbf{y^p}) - \ell(\mathbf{p_z}, \mathbf{y^{p_z}})]| \lesssim \sigma \tag{6}$$

This is an application of Lemma 21 to the bounded function $\ell(\mathbf{p_z}, y)$, and then sampling y according to $\mathbf{y^p}$ and $\mathbf{y^{p_z}}$, and then taking expectations over $\mathbf{p_z}$.

Equation (4) follows from Equations (5) and (6) and the triangle inequality. ◀

Proof of Theorem 8. Lemma 15, implies the existence of a joint distribution $(\mathbf{p}, \mathbf{q}, \mathbf{y})$ of random variables in $[0,1] \times [0,1] \times \{0,1\}$ such that $(\mathbf{p}, \mathbf{y}) \sim \mu$, $(\mathbf{q}, \mathbf{y}) \sim \nu$ and $\mathbf{E}[|\mathbf{p} - \mathbf{q}|] = W(\mu, \nu)$.

Let $\mathbf{z} \in [\pm\sigma]$ for $\sigma \in (0,1]$. Recall that $\mathbf{y^{p_z}} \in \{0,1\}$ is such that $\mathbf{E}[\mathbf{y^{p_z}}|\mathbf{p_z}] = \mathbf{p_z}$, in other words, $\mathbf{p_z}$ is the Bayes optimal predictor for $\mathbf{y^{p_z}}$. Since ℓ is proper, for any $\kappa : [0,1] \to [0,1]$:

$$\mathbf{E}[\ell(\mathbf{p_z}, \mathbf{y^{p_z}})] \le \mathbf{E}[\ell(\kappa(\mathbf{p_z}), \mathbf{y^{p_z}})]. \tag{7}$$

By applying Lemma 23 to the loss function ℓ,

$$|\mathbf{E}[\ell(\mathbf{p_z}, \mathbf{y})] - \mathbf{E}[\ell(\mathbf{p_z}, \mathbf{y^{p_z}})]| \lesssim \sigma + \frac{\mathsf{smCE}(\mu)}{\sigma}. \tag{8}$$

Since Lemma 23 holds for any bounded loss, even if it is not proper, we may apply it for $\tilde{\ell}(p,y) = \ell(\kappa(p), y)$ which is clearly bounded. We obtain the following:

$$|\mathbf{E}[\ell(\kappa(\mathbf{p_z}), \mathbf{y^{p_z}})] - \mathbf{E}[\ell(\kappa(\mathbf{p_z}), \mathbf{y})]| \lesssim \sigma + \frac{\mathsf{smCE}(\mu)}{\sigma}. \tag{9}$$

By Lemma 22, for $y \in \{0,1\}$, the function

$$f_y(p) = \mathop{\mathbf{E}}_{\mathbf{z} \sim [\pm\sigma]}[\ell(\kappa(p_{\mathbf{z}}), y)]$$

is $O(1/\sigma)$-Lipschitz in p, hence $|f_y(p) - f_y(q)| \le O(|p-q|/\sigma)$. Taking expectations over $\mathbf{p}, \mathbf{q}$ and $\mathbf{y}$, we get:

$$\left|\mathbf{E}[\ell(\kappa(\mathbf{p_z}), \mathbf{y})] - \mathbf{E}[\ell(\kappa(\mathbf{q_z}), \mathbf{y})]\right| \lesssim \frac{\mathbf{E}[|\mathbf{p} - \mathbf{q}|]}{\sigma}. \tag{10}$$

Using Equations (9) and (10), the triangle inequality, and the fact that $\mathbf{E}[|\mathbf{p} - \mathbf{q}|] = W(\mu, \nu)$

$$\left|\mathbf{E}[\ell(\kappa(\mathbf{p_z}), \mathbf{y^{p_z}})] - \mathbf{E}[\ell(\kappa(\mathbf{q_z}), \mathbf{y})]\right| \lesssim \sigma + \frac{\mathsf{smCE}(\mu)}{\sigma} + \frac{W(\mu,\nu)}{\sigma}. \tag{11}$$

The desired result follows by starting from Equation (7) and

- Replacing $\mathbf{E}[\ell(\mathbf{p_z}, \mathbf{y^{p_z}})]$ on the LHS with $\mathbf{E}[\ell(\mathbf{p_z}, \mathbf{y})]$ by Equation (8)
- Replacing $\mathbf{E}[\ell(\kappa(\mathbf{p_z}), \mathbf{y^{p_z}})]$ on the RHS with $\mathbf{E}[\ell(\kappa(\mathbf{q_z}), \mathbf{y})]$ by Equation (11). ◀

4 Smooth Calibration Error Approximates Lower Distance

In this section, we prove Lemmas 12 and 13, which immediately imply Theorem 4 when combined. To begin, we recall their respective statements:

▶ **Lemma 12.** $1/2 \le \mathsf{dEMC}(\mu)/\mathsf{smCE}(\mu) \le 2$ *for every PLD* μ.

▶ **Lemma 13.** $\underline{\mathsf{dCE}}(\mu) = \mathsf{dEMC}(\mu)$ *for every PLD* μ.

Of the two proofs, the proof of Lemma 12 is significantly simpler:

Proof of Lemma 12. To prove the lower bound, we must show that $\mathsf{smCE}(\mu) < 2\varepsilon$ whenever $\mathsf{dEMC}(\mu) < \varepsilon$. First, by the definition of dEMC, we can construct jointly distributed pairs $(\mathbf{p}, \mathbf{y})$ and $(\mathbf{p}', \mathbf{y}')$ such that the former is distributed according to μ, the latter is perfectly calibrated, and the two are close in expectation:

$$\mathbf{E}\left[d\big((\mathbf{p}, \mathbf{y}), (\mathbf{p}', \mathbf{y}')\big)\right] = \mathbf{E}|\mathbf{p} - \mathbf{p}'| + \mathbf{E}|\mathbf{y} - \mathbf{y}'| < \varepsilon.$$

Next, to bound $\mathsf{smCE}(\mu)$, consider any 1-Lipschitz function $w : [0, 1] \to [-1, 1]$. Then,

$$\begin{aligned}&\mathbf{E}\big[w(\mathbf{p})(\mathbf{y} - \mathbf{p})\big] = \\ &= \underbrace{\mathbf{E}\big[w(\mathbf{p})\big((\mathbf{y} - \mathbf{p}) - (\mathbf{y}' - \mathbf{p}')\big)\big]}_{<\varepsilon} + \underbrace{\mathbf{E}\big[\big(w(\mathbf{p}) - w(\mathbf{p}')\big)(\mathbf{y}' - \mathbf{p}')\big]}_{<\varepsilon} + \underbrace{\mathbf{E}\big[w(\mathbf{p}')(\mathbf{y}' - \mathbf{p}')\big]}_{=0}.\end{aligned}$$

In the above equation, the first two terms on the right are $< \varepsilon$ by the ε-closeness of $(\mathbf{p}, \mathbf{y})$ and $(\mathbf{p}', \mathbf{y}')$ in expectation. The third term is 0 by calibration of $(\mathbf{p}', \mathbf{y}')$, which means that $\mathbf{E}[\mathbf{y}'|\mathbf{p}'] = \mathbf{p}'$.

To prove the upper bound, we use the observation from [6] that when the smooth calibration error is small, it is easy to construct a nearby perfectly calibrated predictor by altering the conditional distribution of the label. Specifically, if $(\mathbf{p}, \mathbf{y}) \sim \mu$, then define $\tilde{\mathbf{y}}|\mathbf{p} \sim \mathrm{Ber}(\mathbf{p})$ and let ν be the distribution of $(\mathbf{p}, \tilde{\mathbf{y}})$. Then, $\nu \in \mathsf{Cal}$ and for any function $f : \mathcal{S} \to [-1, +1]$, we have

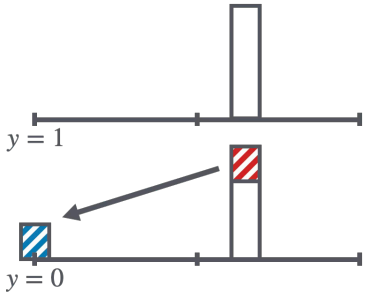
$$\big|\mathbf{E}[f(\mathbf{p}, \mathbf{y})] - \mathbf{E}[f(\mathbf{p}, \tilde{\mathbf{y}})]\big| = \Big|\mathbf{E}\big[(f(\mathbf{p}, 1) - f(\mathbf{p}, 0))(\mathbf{y} - \mathbf{p})\big]\Big|.$$

By Kantorovich-Rubinstein duality, the supremum of the left side over all 1-Lipschitz f is equal to $W(\mu, \nu)$. The right side is at most $2\mathsf{smCE}(\mu)$ since the Lipschitz constant of $f(\cdot, 1) - f(\cdot, 0)$ is at most twice that of f. We conclude that $\mathsf{dEMC}(\mu) \le 2\mathsf{smCE}(\mu)$. ◀

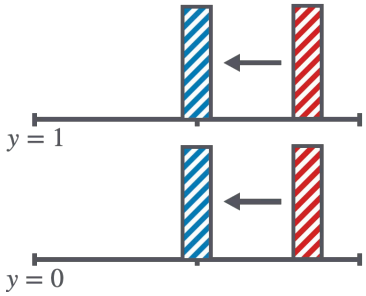
Next, we prove Lemma 13, which equates dEMC and $\underline{\mathsf{dCE}}$ via a novel but simple exchange argument. In what follows, a *transport plan* π from μ to Cal is simply a coupling of μ with some $\nu \in \mathsf{Cal}$. When we say that a plan π *moves* mass m from $A \subseteq \mathcal{S}$ to $B \subseteq \mathcal{S}$, we mean that the associated coupling assigns mass m to the set $A \times B$. Similarly, the *cost* of a transport plan π is

$$\mathop{\mathbf{E}}_{((\mathbf{p}, \mathbf{y}), (\mathbf{p}', \mathbf{y}')) \sim \pi}\Big[d\big((\mathbf{p}, \mathbf{y}), (\mathbf{p}', \mathbf{y}')\big)\Big].$$

Proof of Lemma 13. Let $S_0 = [0, 1] \times \{0\}$ and $S_1 = [0, 1] \times \{1\}$. As subsets of the domain $\mathcal{S}$, these line segments correspond to the events $\mathbf{y} = 0$ and $\mathbf{y} = 1$. By definition, $\mathsf{dEMC}(\mu)$ is the infimum cost of all transport plans from μ to Cal, and $\underline{\mathsf{dCE}}(\mu)$ is the infimum cost of transport plans from μ to Cal that move no mass between S_0 and S_1. In more detail: We know from



(a) Move some mass to corner.



(b) Move all mass where ratio dictates.

Figure 2 Visualization of the proof of Lemma 13.

Lemma 14 that $\underline{\mathsf{dCE}}$ is the infimum cost over triples $(\mathbf{p}, \mathbf{q}, \mathbf{y})$, as opposed to generic couplings of $(\mathbf{p}, \mathbf{y})$ with $(\mathbf{q}, \mathbf{y}')$ for some $\mathbf{y}'$. This precisely means enforcing $\Pr[\mathbf{y} \neq \mathbf{y}'] = 0$ under the coupling, a.k.a. no mass is moved between S_0 and S_1. (Recall that just before the proof started, we clarified that "moving mass from A to B" means the coupling placing mass on $A \times B$.)

Thus, to prove that $\mathsf{dEMC}(\mu) = \underline{\mathsf{dCE}}(\mu)$, it suffices to show that for every transport plan π from μ to Cal, there exists a transport plan π' from μ to Cal that moves no mass between S_0 and S_1 and is no costlier than π. By taking limits, it suffices to prove the claim in the case that all distributions under consideration are discrete.

To prove the claim, we first observe that π can be viewed as the composition of two consecutive plans π_1 and π_2 such that π_1 only moves mass within each segment and π_2 only moves mass between $(p, 0)$ and $(p, 1)$ for various values $p \in [0, 1]$. Visually, π_1 moves mass "horizontally," and π_2 moves mass "vertically," as depicted in Figure 2.

Next, suppose that π_2 moves some mass from $(p, 0)$ to $(p, 1)$ for some value $p \in [0, 1]$. If m_0 and m_1 were the masses at these points just before the move, then π_2 must have moved precisely $c = pm_0 - (1-p)m_1$ mass between these two points to achieve calibration. Indeed, after such a move, which costs precisely c, the two points would have masses $m_0 - c = (1-p)(m_0 + m_1)$ and $m_1 + c = p(m_0 + m_1)$, respectively, which are in the correct ratio for calibration.

Alternatively, as shown in Figure 2a, we could have achieved calibration by moving $m' = m_0 - (1/p - 1)m_1$ mass from $(p, 0)$ to $(0, 0)$, which does not cross segments and has the same total cost of $pm' = c$. Indeed, after such a move, the two points would have masses $(1/p - 1)m_1$ and m_1, respectively, which are in the correct ratio for calibration. (Yet another option, shown in Figure 2b, would be to move all the mass at $(p, 0)$ and $(p, 1)$ to the points $(p', 0)$ and $(p', 1)$, respectively, where $p' = m_1/(m_0 + m_1)$. This also has total cost $(m_0 + m_1)(p - p') = c$ and satisfies calibration.)

At this point, we have shown that any plan π can be transformed into a plan π' with the same cost that moves no mass from S_0 to S_1. A similar argument applies to shipments from S_1 to S_0. Thus, there exists an optimal transport plan that moves no mass between segments in *either* direction, as claimed. ◀

Next, we prove Corollary 17. We recall the statement here for convenience.

▶ **Corollary 17.** *Given a PLD μ, the following are equal up to constant factors:*

(a) *Its earth mover's distance to calibration,* $\mathsf{dEMC}(\mu)$*:*

$$\inf\left\{W(\mu,\nu)\,\middle|\,\nu\in\mathsf{Cal}\right\},$$

(b) *Its earth mover's distance to calibration while preserving the marginal of* $\mathbf{y}$*:*

$$\inf\left\{W(\mu,\nu)\,\middle|\,\nu\in\mathsf{Cal}^{=\tau}\right\},$$

where $\tau=\Pr_{(\mathbf{p},\mathbf{y})\sim\mu}[\mathbf{y}=1]$.

(c) *Its earth mover's distance to calibration while preserving the marginal of* $\mathbf{p}$*:*

$$\inf\left\{W(\mu,\nu)\,\middle|\,\nu\in\mathsf{Cal}\text{ and }\Pr_{(\mathbf{p},\mathbf{y})\sim\mu}[\mathbf{p}\le t]=\Pr_{(\mathbf{p},\mathbf{y})\sim\nu}[\mathbf{p}\le t]\text{ for all }t\in[0,1]\right\}.$$

More precisely, $a=b\le c\le 2a$.

Proof. The inequalities $a\le b$, $a\le c$, and $b\le \underline{\mathrm{dCE}}(\mu)$ are all clear, since each inequality compares the infimum over a set to the infimum over a subset. Lemma 13 states that $\underline{\mathrm{dCE}}(\mu)=a$. Finally, our proof of the upper bound in Lemma 12 implies that $c\le 2a$ since the only calibrated prediction-label distribution which preserves the marginal of $\mathbf{p}$ is that of $(\mathbf{p},\tilde{\mathbf{y}})$, where $\tilde{\mathbf{y}}|\mathbf{p}\sim\mathrm{Bernoulli}(\mathbf{p})$. ◀

Finally, we prove Lemma 15. The proof is similar in spirit to the proof of Lemma 13 but simpler. Again, we recall the relevant statement for convenience.

▶ **Lemma 15.** *For any two $\mu,\nu\in\mathsf{PId}^{=\tau}$ for $\tau\in[0,1]$, we have that $W(\mu,\nu)=\inf\mathbf{E}[|\mathbf{p}-\mathbf{q}|]$, where the infimum is taken over all triples of joint random variables $(\mathbf{p},\mathbf{q},\mathbf{y})$ such that $(\mathbf{p},\mathbf{y})\sim\mu$ and $(\mathbf{q},\mathbf{y})\sim\nu$.*

Proof. As in the proof of Lemma 13, consider two PLDs μ and ν. In this case, however, we do not assume that either μ or ν is calibrated, but instead require that $\Pr_{(\mathbf{p},\mathbf{y})\sim\mu}[\mathbf{y}=1]=\Pr_{(\mathbf{p},\mathbf{y})\sim\nu}[\mathbf{y}=1]=\tau$. Phrased differently, $\mu,\nu\in\mathsf{PId}^{=\tau}$ for the same value $\tau\in[0,1]$.

Recall that $W(\mu,\nu)$ is, by definition, the infimum cost among all transport plans π from μ to ν (i.e. couplings of μ and ν), which may or may not move mass between the segments S_0 and S_1. Observe that in the case that π *does* move mass from S_0 to S_1, it must also move the same amount of mass from S_1 back to S_0, from our assumption that both μ and ν place exactly τ mass on S_1. Similarly, if π moves mass from S_1 to S_0, it must move the same amount of mass from S_0 back to S_1.

Ultimately, what we want to prove is the following claim: $W(\mu,\nu)$ is in fact equal to the infimum cost among the *restricted* set of couplings of $(\mathbf{p},\mathbf{y})\sim\mu$ with $(\mathbf{q},\mathbf{y}')\sim\nu$ such that $\Pr[\mathbf{y}=\mathbf{y}']=1$, a.k.a. triples $(\mathbf{p},\mathbf{q},\mathbf{y})$ with $(\mathbf{p},\mathbf{y})\sim\mu$ and $(\mathbf{q},\mathbf{y})\sim\nu$. Viewing the coupling as a transport plan, this is equivalent to the restriction that the plan π moves *no* mass from S_0 to S_1 and *no* mass from S_1 to S_0.

To prove the claim, let π be any transport plan from $\mu,\nu\in\mathsf{PId}^{=\tau}$, and factor π into two consecutive plans π_1 and π_2 as in the proof of Lemma 13 and as depicted in Figure 2, where π_1 only moves mass within segments ("horizontally"), and π_2 only moves mass between pairs of points of the form $(p,0)$ and $(p,1)$ ("vertically").

Suppose for the sake of contradiction that π_2 moved some positive mass $c > 0$ from S_0 to S_1 or vice versa. By the preceding discussion, we know that π_2 must in fact move c mass from *both* S_0 to S_1 and from S_1 to S_0. Recall that the metric under consideration is ℓ_1:

$$d\big((p, y), (p', y')\big) = |p - p'| + |y - y'|.$$

Therefore, the cost of π_2 is $2c$. Had we instead exchanged these two shipments and moved them to their final destinations within segments, the above equality shows that we would have incurred a cost of $\leq 2c$, proving the claim. ◀

References

1. Jaroslaw Blasiok, 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, STOC 2023*, pages 1727–1740. ACM, 2023. doi:10.1145/3564246.3585182.
2. Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 72071–72095. Curran Associates, Inc., 2023.
3. Jaroslaw Blasiok and Preetum Nakkiran. Smooth ECE: principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024.
4. A. P. Dawid. Objective probability forecasts. *University College London, Dept. of Statistical Science. Research Report 14*, 1982.
5. Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.
6. Parikshit Gopalan and Lunjia Hu. Calibration through the lens of indistinguishability. *ACM SIGecom Exchanges*, 23(1), July 2025. URL: https://www.sigecom.org/exchanges/volume_23/1/HU.pdf.
7. Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *Innovations in theoretical computer science (ITCS'23)*, 2023.
8. Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *Innovations in Theoretical Computer Science (ITCS'2022)*, 2022. arXiv:2109.05389.
9. Parikshit Gopalan, Konstantinos Stavropoulos, Kunal Talwar, and Pranay Tankala. Efficient calibration for decision making. In *Proceedings of the ACM SIGACT Symposium on Theory of Computing (STOC'2026), to appear*, 2026.
10. Jason D. Hartline, Yifan Wu, and Yunran Yang. Smooth calibration and decision making. In *6th Symposium on Foundations of Responsible Computing, FORC 2025*, volume 329 of *LIPIcs*, pages 16:1–16:26, 2025. doi:10.4230/LIPIcs.FORC.2025.16.
11. Lunjia Hu and Yifan Wu. Calibration error for decision making. In *Proceedings of the 65th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2024.
12. Sham Kakade and Dean Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008. doi:10.1016/J.JCSS.2007.04.017.
13. Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5143–5145. PMLR, 12–15 July 2023. URL: https://proceedings.mlr.press/v195/kleinberg23a.html.
14. Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2805–2814. PMLR, 2018.

A Omniprediction With Respect to Nearby Calibrated Predictors

In this section, we prove Theorem 10, where we restrict the baselines class to calibrated predictors in $\mathsf{Cal}^{=\tau(\mathcal{D})}$. We restate the theorem below.

▶ **Theorem 10.** *Let $p \in \mathsf{P}(\mathcal{X})$, $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{0,1\}$, $\mu = p \circ \mathcal{D}$ and $\nu \in \mathsf{Cal}^{=\tau(\mathcal{D})}$. For any bounded proper loss $\ell \in \mathcal{L}$, and $\sigma \in (0,1]$*

$$\mathop{\mathbf{E}}_{\substack{(\mathbf{p},\mathbf{y})\sim p\circ\mathcal{D}\\ \mathbf{z}\sim[\pm\sigma]}}[\ell(\mathbf{p_z},\mathbf{y})] - \mathop{\mathbf{E}}_{(\mathbf{q},\mathbf{y})\sim\nu}[\ell(\mathbf{q},\mathbf{y})] \lesssim \sigma + \frac{\mathsf{smCE}(\mu)}{\sigma} + W(\mu,\nu).$$

Our proof will use the V-shaped losses introduced in [13]. These are losses of the form

$$\ell_v(p,y) = -(y-v)\mathrm{sign}(p-v),\ v \in [0,1].$$

It is easy to see that $\ell_v \in \mathcal{L}$. In fact, these functions form a basis for $\mathcal{L}$: [13] show that every loss $\ell \in \mathcal{L}$ can be written as

$$\ell(p,q) = \sum_v \lambda_v \ell_v(p,q), \quad \lambda_v \geq 0, \sum_v \lambda_v \leq 2. \tag{12}$$

For a precise statement that accounts for convergence issues, see [9, Lemma 3.5]. This statement ignores linear terms of the form $ay+b$ for $a,b \in \mathbb{R}$. Since the contribution of such terms is independent of the prediction p, we can ignore them as long as we compare PLDs which have the same marginal distribution over $\mathbf{y}$.

For a loss ℓ_v and a perfectly calibrated PLD $\mu \in \mathsf{Cal}$, we can compute the expected loss as follows:

$$\mathop{\mathbf{E}}_{(\mathbf{p},\mathbf{y})\sim\mu}[\ell_v(\mathbf{p},\mathbf{y})] = \mathbf{E}[-(\mathbf{y}-v)\mathrm{sign}(\mathbf{p}-v)] = \mathbf{E}[-(\mathbf{p}-v)\mathrm{sign}(\mathbf{p}-v)] = -\mathbf{E}[|\mathbf{p}-v|] \tag{13}$$

Next we provide some helpful lemmas whose proofs can be found in the full version (see "related version"). The first shows that when we restrict our attention to fully calibrated predictors, the smoothing operation does not significantly change the value of any proper loss.

▶ **Lemma 24.** *Let $(\mathbf{q}^*,\mathbf{y}) \in \mathsf{Cal}$ and let $\mathbf{z} \sim [\pm\sigma]$. Then for any proper loss $\ell \in \mathcal{L}$:*

$$\left|\mathbf{E}[\ell(\mathbf{q}^*,\mathbf{y})] - \mathbf{E}[\ell(\mathbf{q}^*_\mathbf{z},\mathbf{y})]\right| \lesssim \sigma.$$

Our next lemma shows that when restricted to the space Cal of calibrated PLDs, every loss in $\mathcal{L}$ is Lipschitz in the earth mover distance between them.

▶ **Lemma 25.** *Let $\mu,\nu \in \mathsf{Cal}^{=\tau}$. Then, for any $\ell \in \mathcal{L}$,*

$$\left|\mathop{\mathbf{E}}_{(\mathbf{q},\mathbf{y})\sim\mu}[\ell(\mathbf{q},\mathbf{y})] - \mathop{\mathbf{E}}_{(\mathbf{q}',\mathbf{y}')\sim\nu}[\ell(\mathbf{q}',\mathbf{y}')]\right| \leq W(\mu,\nu).$$

We are now ready to prove Theorem 10.

Proof of Theorem 10. Consider the joint distribution $(\mathbf{p},\mathbf{q}^*,\mathbf{y})$ such that $(\mathbf{q}^*,\mathbf{y}) \sim \nu^*$ for $\nu^* \in \mathsf{Cal}^{=\tau}$, $(\mathbf{p},\mathbf{y}) \sim \mu$ and $\mathbf{E}[|\mathbf{p}-\mathbf{q}^*|]$ is the minimum possible. By Lemma 14 and Corollary 16, this minimum is $\underline{\mathsf{dCE}}(\mu) = W(\mu,\nu^*)$. Note that the distribution ν^* of $(\mathbf{q}^*,\mathbf{y})$ need not be the same as ν, it could be that $W(\mu,\nu^*) \leq W(\mu,\nu)$. We invoke Theorem 8 with κ being the identity function to obtain the following:

$$\mathbf{E}[\ell(\mathbf{p_z},\mathbf{y})] - \mathbf{E}[\ell(\mathbf{q}^*_\mathbf{z},\mathbf{y})] \lesssim \sigma + \frac{1}{\sigma}(\mathsf{smCE}(\mu) + E[|\mathbf{p}-\mathbf{q}^*|]) \lesssim \sigma + \frac{\mathsf{smCE}(\mu)}{\sigma}. \tag{14}$$

By Lemma 24, we get:

$$\mathbf{E}[\ell(\mathbf{q}^*_{\mathbf{z}}, \mathbf{y})] - \mathbf{E}[\ell(\mathbf{q}^*, \mathbf{y})] \lesssim \sigma.$$

By Lemma 25 applied to the calibrated PLDs ν^* and ν,

$$\begin{aligned}|\mathop{\mathbf{E}}_{(\mathbf{q}^*,\mathbf{y})\sim\nu^*}[\ell(\mathbf{q}^*, \mathbf{y})] - \mathop{\mathbf{E}}_{(\mathbf{q},\mathbf{y})\sim\nu}[\ell(\mathbf{q}, \mathbf{y})]| &\le W(\mathbf{q}^*, \mathbf{q}) \le W(\nu^*, \nu) \le W(\mu, \nu^*) + W(\mu, \nu) \\ &\le 2\mathsf{smCE}(\mu) + W(\mu, \nu).\end{aligned}$$

We obtain the desired result by replacing $\ell(\mathbf{q}^*_{\mathbf{z}}, \mathbf{y})$ on the LHS of Equation (14) with $\ell(\mathbf{q}, \mathbf{y}^{\mathbf{q}})$, at a further cost of $O(\sigma + \mathsf{smCE}(\mu)) + W(\mu, \nu)$. ◀

B Inapproximability of Upper Distance

In this section, we state and prove the full version of Theorem 20.

▶ **Theorem 26.** *Fix $\varepsilon, \varepsilon' > 0$, $k \in \mathbb{N}$, and $\delta_{0j}, \delta_{1j} \in \mathbb{R}$ for $j \in [k]$. Consider the following tuples $(\mathcal{X}, \mathcal{D}_{\mathcal{X}}, p^*, p)$:*

(a) *Let $\mathcal{X}_a = \{L, R\}$ and define $\mathcal{D}_{\mathcal{X}_a}$, p^*_a, and p_a as follows:*

x	$\mathcal{D}_{\mathcal{X}_a}(x)$	$p^*_a(x)$	$p_a(x)$
L	$1/2$	$1/2$	$1/2 - \varepsilon$
R	$1/2$	$1/2$	$1/2 + \varepsilon$

(b) *Let $\mathcal{X}_b = \{L_0, L_1, R_0, R_1\}$ and define $\mathcal{D}_{\mathcal{X}_b}$, p^*_b, and p_b as follows:*

x	$\mathcal{D}_{\mathcal{X}_b}(x)$	$p^*_b(x)$	$p_b(x)$
L_0	ε'	1	$1/2 - \varepsilon$
L_1	$1/2 - \varepsilon'$	$1/2 - \varepsilon$	$1/2 - \varepsilon$
R_0	ε'	0	$1/2 + \varepsilon$
R_1	$1/2 - \varepsilon'$	$1/2 + \varepsilon$	$1/2 + \varepsilon$

(c) *Building on (a), let $\mathcal{X}_c = \{L, R\} \times [k/2]$ and define $\mathcal{D}_{\mathcal{X}_c}$, p^*_c, and p_c as follows:*

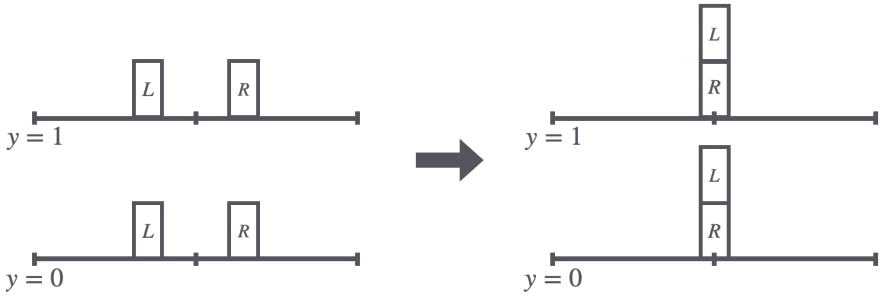
x	$\mathcal{D}_{\mathcal{X}_c}(x)$	$p^*_c(x)$	$p_c(x)$
(L, j)	$1/k$	$1/2$	$1/2 - \varepsilon + \delta_{0j}$
(R, j)	$1/k$	$1/2$	$1/2 + \varepsilon + \delta_{1j}$

(d) *Building on (b), let*

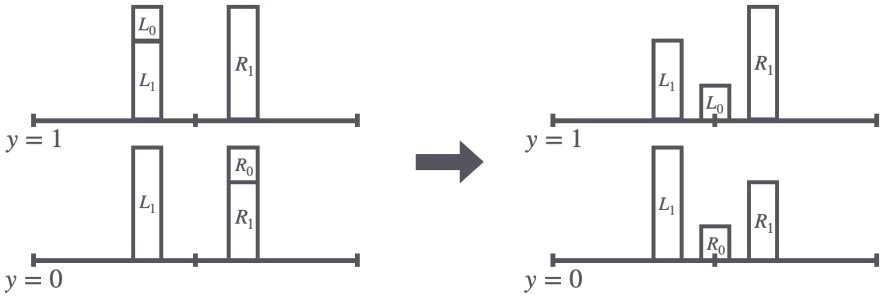
$$\mathcal{X}_d = \big(\{L_0, R_0\} \times \{1, \ldots, \varepsilon' k\}\big) \cup \big(\{L_1, R_1\} \times \{\varepsilon' k + 1, \ldots, k/2\}\big)$$

and define $\mathcal{D}_{\mathcal{X}_d}$, p^*_d, and p_d as follows:

x	$\mathcal{D}_{\mathcal{X}_d}(x)$	$p^*_d(x)$	$p_d(x)$
(L_0, j)	$1/k$	1	$1/2 - \varepsilon + \delta_{0j}$
(L_1, j)	$1/k$	$1/2 - \varepsilon$	$1/2 - \varepsilon + \delta_{0j}$
(R_0, j)	$1/k$	0	$1/2 + \varepsilon + \delta_{1j}$
(R_1, j)	$1/k$	$1/2 + \varepsilon$	$1/2 + \varepsilon + \delta_{1j}$



(a) Cases (a) and (c) of Theorem 26.



(b) Cases (b) and (d) of Theorem 26.

Figure 3 Optimal transport to calibration for each case of Theorem 26. In cases (a) and (b), the predictions are concentrated entirely on the points $1/2 \pm \varepsilon$, but in cases (c) and (d), they are instead scattered nearby.

Suppose that $\varepsilon' = \varepsilon/(1+2\varepsilon)$, that $\varepsilon' k$ and $k/2$ are both integers, and that the parameters δ_{ij} are sampled i.i.d. from a continuous distribution over $[-\varepsilon/2, \varepsilon/2]$. Then,

- **(i)** *Case (a), which has $\mathsf{dCE} \geq \Omega(\varepsilon)$, is impossible to distinguish with any positive advantage from case (b), which has $\mathsf{dCE} \leq O(\varepsilon^2)$, given any finite number of prediction-label samples.*
- **(ii)** *It requires at least $\Omega(\sqrt{k})$ prediction-label samples to distinguish (with constant advantage in expectation over the choice of δ_{ij}) case (c), which has $\overline{\mathsf{dCE}} \geq \Omega(\varepsilon)$, from case (d), which has $\overline{\mathsf{dCE}} \leq O(\varepsilon^2)$ (with probability 1 over the choice of δ_{ij}).*

In particular, dCE cannot be estimated from prediction-label samples within a better-than-quadratic factor, and $\overline{\mathsf{dCE}}$ cannot be estimated within a better-than-quadratic factor from any number of prediction-label samples that is independent of the support size of the distribution of predictions.

Observe that part (i) of Theorem 26 is precisely the prior result of [1] regarding the inapproximability of dCE in the prediction-only access model. Part (ii) is our new result regarding $\overline{\mathsf{dCE}}$. The proofs of both parts are most easily illustrated by Figure 3.

Proof of Theorem 26.

- **(i)** To see that case (a) has $\mathsf{dCE} \geq \Omega(\varepsilon)$, observe that the Bayes optimal predictor p^*_a is the constant $1/2$ function. Therefore, the only calibrated predictor on the domain $\mathcal{X}_a$ is the constant $1/2$ function itself. Since p_a always outputs values exactly ε-far away from $1/2$, it follows that $\mathsf{dCE} = \varepsilon$ in case (a). In Figure 3a, this corresponds to the movement of all mass (total of 1) across a distance of length ε, for a total cost of $1 \cdot \varepsilon \leq O(\varepsilon)$.

In contrast, the fact that case (b) has $\mathsf{dCE} \leq O(\varepsilon^2)$ is witnessed by calibrated predictor

$$p'_b(x) = \begin{cases} 1/2 & \text{if } x \in \{L_0, R_0\} \\ 1/2 - \varepsilon & \text{if } x = L_1, \\ 1/2 + \varepsilon & \text{if } x = R_1. \end{cases}$$

Visually, p'_b is obtained from p_b by "moving" the two ε' masses at $x \in \{L_0, R_0\}$, each over a distance of length ε, to the point $1/2$, as shown in Figure 3b, for a total cost of $2\varepsilon' \cdot \varepsilon \leq O(\varepsilon^2)$. More formally,

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}_b}} \left| p_b(\mathbf{x}) - p'_b(\mathbf{x}) \right| = 2\varepsilon' \cdot \varepsilon + (1 - 2\varepsilon') \cdot 0 = \frac{2\varepsilon^2}{1 + 2\varepsilon} \leq O(\varepsilon^2).$$

At this point, all that remains is to show that cases (a) and (b) give rise to identical PLDs. This will imply that they cannot be distinguished from any finite number of prediction-label samples, as claimed. For this, we simply observe that in case (b), the expected label $\mathbf{y}$ given $p_b(\mathbf{x}) = 1/2 - \varepsilon$ is

$$\mathbf{E}\Big[\mathbf{y} \,\Big|\, p_b(\mathbf{x}) = \frac{1}{2} - \varepsilon\Big] = \frac{\varepsilon' \cdot 1 + \left(\frac{1}{2} - \varepsilon'\right) \cdot \left(\frac{1}{2} - \varepsilon\right)}{\varepsilon' + \left(\frac{1}{2} - \varepsilon'\right)} = \frac{1}{2}.$$

Similarly, one can check that $\mathbf{E}[\mathbf{y} | p_b(\mathbf{x}) = 1/2 + \varepsilon] = 1/2$. Therefore, cases (a) and (b) give rise to the same PLDs, as claimed, completing the proof of part (i).

(ii) First, note that cases (c) and (d) are generalizations of cases (a) and (b), respectively, where instead of predicting $1/2 \pm \varepsilon$, we make predictions that are small random perturbations of $1/2 \pm \varepsilon$. Thus, our proof strategy for part (ii) will be similar in spirit to our strategy for part (i).

To see that case (c) has $\overline{\mathsf{dCE}} \geq \Omega(\varepsilon)$, observe that the Bayes optimal predictor p^*_c is the constant $1/2$ function. Therefore, the only calibrated predictor on the domain $\mathcal{X}_c$ is the constant $1/2$ function itself. Since $|\delta_{ij}| \leq \varepsilon/2$, the predictor p_c always outputs values at least $\varepsilon/2$-far away from $1/2$. It follows that $\overline{\mathsf{dCE}} \geq \mathsf{dCE} \geq \varepsilon/2$ in case (c) (with probability 1 over the choice of δ_{ij}). We also have $\overline{\mathsf{dCE}} \leq 3\varepsilon/2$, and this is again depicted in Figure 3a.

The situation in case (d) is somewhat subtler. Intuitively, in case (d), there always *exists* a "cheap" post-processing κ certifying $\overline{\mathsf{dCE}} \leq O(\varepsilon^2)$, but such a κ is hard to actually *find* from a handful of prediction-label samples. This is in contrast to parts (a) and (c), where such a post-processing does *not* exist, and part (b), where such a post-processing exists *and* is easy to construct from samples.

More formally, the fact that case (d) has $\overline{\mathsf{dCE}} \leq O(\varepsilon^2)$ is witnessed by the following post-processing function κ_d. The following definition of κ_d mimics the transition from the predictor p_b to p'_b that we used in case (b) and visualized in Figure 3b. In defining κ_d, we crucially use the fact that all perturbation terms δ_{ij} are distinct. This occurs with probability 1 since they are drawn i.i.d. from a continuous distribution:

$$\kappa_d(p) = \begin{cases} 1/2 & \text{if } p = 1/2 - \varepsilon + \delta_{0j} \text{ for some } j \leq \varepsilon' k, \\ 1/2 - \varepsilon & \text{if } p = 1/2 - \varepsilon + \delta_{0j} \text{ for some } j > \varepsilon' k, \\ 1/2 & \text{if } p = 1/2 + \varepsilon + \delta_{1j} \text{ for some } j \leq \varepsilon' k, \\ 1/2 + \varepsilon & \text{if } p = 1/2 + \varepsilon + \delta_{1j} \text{ for some } j > \varepsilon' k. \end{cases}$$

Indeed, essentially the same analysis as in part (b) shows that $\kappa \circ p_d$ is calibrated and

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}_d}} \left|\kappa(p_d(\mathbf{x})) - p_d(\mathbf{x}))\right| \leq O(\varepsilon^2).$$

Finally, we argue that cases (c) and (d) are hard to distinguish from few i.i.d. prediction-label samples $(p(\mathbf{x}_1), \mathbf{y}_1), \ldots, (p(\mathbf{x}_s), \mathbf{y}_s)$. For this, observe that in a sample of size s, the probability that we see a collision (i.e. sample $(p(\mathbf{x}), \mathbf{y})$ for the same point $\mathbf{x} = x \in \mathcal{X}$ twice) is at most $\binom{s}{2}/k \leq O(s^2/k)$. Moreover, by construction, conditional on seeing no collisions, each conditional distribution $\mathbf{y}_i | p(\mathbf{x}_i)$ is uniform on $\{0, 1\}$ in both cases (c) and (d) over the randomness in the choice of the parameters δ_{ij}. This is clear for case (c), whereas for case (d), it follows from the same calculation as in part (b):

$$\frac{\varepsilon' \cdot 1 + \left(\frac{1}{2} - \varepsilon'\right) \cdot \left(\frac{1}{2} - \varepsilon\right)}{\varepsilon' + \left(\frac{1}{2} - \varepsilon'\right)} = \frac{1}{2}.$$

We conclude that to distinguish cases (c) and (d) with at least constant advantage over the randomness in δ_{ij}, we require at least $s \geq \Omega(\sqrt{k})$ prediction-label samples. ◀