

# Efficient Calibration for Decision Making

Parikshit Gopalan  
Apple

Konstantinos Stavropoulos\*  
UT Austin

Kunal Talwar  
Apple

Pranay Tankala\*  
Harvard

## Abstract

A decision-theoretic characterization of perfect calibration is that an agent seeking to minimize a proper loss in expectation cannot improve their outcome by post-processing a perfectly calibrated predictor. Hu and Wu (FOCS'24) use this to define an approximate calibration measure called calibration decision loss (CDL), which measures the maximal improvement achievable by any post-processing over any proper loss. Unfortunately, CDL turns out to be intractable to even weakly approximate in the offline setting, given black-box access to the predictions and labels.

We suggest circumventing this by restricting attention to structured families of post-processing functions  $\mathcal{K}$ . We define the calibration decision loss relative to  $\mathcal{K}$ , denoted  $\text{CDL}_{\mathcal{K}}$  where we consider all proper losses but restrict post-processings to a structured family  $\mathcal{K}$ . We develop a comprehensive theory of when  $\text{CDL}_{\mathcal{K}}$  is information-theoretically and computationally tractable:

- **Complexity characterization.** The sample complexity of estimating  $\text{CDL}_{\mathcal{K}}$  is determined by the VC dimension of  $\text{thr}(\mathcal{K})$ , the concept class consisting of thresholds applied to any  $\kappa \in \mathcal{K}$ . Computationally, estimating  $\text{CDL}_{\mathcal{K}}$  reduces to agnostically learning  $\text{thr}(\mathcal{K})$ . This implies that estimating CDL relative to 1-Lipschitz post-processings is information-theoretically hard.
- **Quantitative characterization.** Augmenting  $\text{thr}(\mathcal{K})$  with indicators of intervals of the form  $[0, a]$  yields a family of weight functions  $\mathcal{K}'$  such that  $\text{CDL}_{\mathcal{K}}$  is characterized, up to a quadratic factor, by the weighted calibration error restricted to  $\mathcal{K}'$ . This significantly generalizes prior bounds that were for specific choices of  $\mathcal{K}$ .
- **Omniprediction.** If  $\text{thr}(\mathcal{K})$  is efficiently learnable there exists a *single* post-processing that performs competitively with the best post-processing in  $\mathcal{K}$  for *every* proper loss. Classical recalibration algorithms including the Pool Adjacent Violators (PAV) algorithm and Uniform-mass binning give similar *omniprediction* guarantees for natural classes of post-processings with monotonic structure.

In addition to introducing new definitions and algorithmic techniques to the theory of calibration for decision making, our results give rigorous guarantees for some widely used recalibration procedures in machine learning.

---

\*Work done during an internship at Apple.

# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Introduction</b>  | <b>1</b>  |
| 1.1      | Approximate Calibration From Indistinguishability.                     | 2         |
| 1.2      | Approximate Calibration From No Regret.                                | 2         |
| <b>2</b> | <b>Overview of Our Results</b>   | <b>5</b>  |
| 2.1      | Definitions  | 5         |
| 2.2      | Testing and Auditing   | 6         |
| 2.3      | Calibration Decision Loss and Weight-Restricted Calibration            | 8         |
| 2.4      | Post-Processing and Omniprediction                                     | 10        |
| <b>3</b> | <b>Preliminaries</b>   | <b>13</b> |
| <b>4</b> | <b>Sample Complexity of Testing and Auditing</b>                       | <b>15</b> |
| 4.1      | A Characterization via VC dimension                                    | 15        |
| 4.2      | Other Families of Loss Functions                                       | 18        |
| <b>5</b> | <b>Relation to Weight-Restricted Calibration</b>                       | <b>19</b> |
| 5.1      | The Characterization   | 20        |
| 5.2      | Translation Invariance Alone Is Insufficient                           | 26        |
| <b>6</b> | <b>Computationally Efficient Testing and Auditing</b>                  | <b>27</b> |
| <b>7</b> | <b>Omniprediction</b>  | <b>29</b> |
| 7.1      | Omniprediction From Agnostic Learning                                  | 29        |
| 7.2      | Pool Adjacent Violators Is an Omnipredictor                            | 31        |
| 7.3      | Omniprediction Through Uniform-Mass Binning and Recalibration          | 33        |
| <b>A</b> | <b>Why Generalized Monotone Functions</b>                              | <b>41</b> |
| <b>B</b> | <b>Additional Proofs</b>   | <b>42</b> |
| B.1      | Additional Proofs From Section 3                                       | 42        |
| B.2      | Additional Proofs From Section 4                                       | 42        |
| <b>C</b> | <b>Tightness of the Weight-Restricted Calibration Characterization</b> | <b>45</b> |
| <b>D</b> | <b>Properties of Generalized Monotone Post-Processings</b>             | <b>46</b> |

# 1 Introduction

Consider the setting of binary classification, where we see examples  $(x, y)$  drawn from a distribution  $\mathcal{D}$  on  $\mathcal{X} \times \{0, 1\}$ . A predictor  $P : \mathcal{X} \rightarrow [0, 1]$  estimates the probability that the label  $y = 1$  for a given  $x$ . While predictions take values in the interval  $[0, 1]$ , the labels are binary. What does it mean for a predictor to be good in such a setting?

The notion of (perfect) calibration, which originates in the forecasting literature [Daw85] requires that every predicted value  $p$  in the range of  $P$ , we have  $\mathbb{E}[y|P(x) = p] = p$ . The Bayes optimal predictor, defined as  $p^*(x) = \mathbb{E}[y|x]$  is calibrated, but calibration is strictly weaker than Bayes optimality. Despite this, calibration gives important guarantees for downstream decision makers who make decisions based on the predictions, where they can trust a calibrated predictor and act *as though it were indeed the Bayes optimal*.

Consider an agent who uses the predictions  $P(x)$  to choose an action  $a \in \mathcal{A}$ , so as to minimize their expected loss  $\mathbb{E}[\ell(a, y)]$  for some loss function  $\ell : \mathcal{A} \times \{0, 1\} \rightarrow \mathbb{R}$ . Such an agent can respond according to the best-response function  $\kappa_\ell : [0, 1] \rightarrow \mathcal{A}$ , where if we denote by  $\text{Ber}(p)$  the Bernoulli distribution with parameter  $p$ , then

$$\kappa_\ell(p) = \arg \min_{a \in \mathcal{A}} \mathbb{E}_{\tilde{y} \sim \text{Ber}(p)} [\ell(a, \tilde{y})].$$

If we wished to minimize loss for the labels  $\tilde{y} \sim \text{Ber}(p)$ , where  $p = P(x)$ , then  $P$  is Bayes optimal. So this corresponds to the agent trusting the predictor  $P$  even for labels  $y$ , as though it were the Bayes optimal predictor.

Calibration gives two important guarantees to any such agent, for every loss function  $\ell$ :

- **No surprises:** The expected loss suffered  $\mathbb{E}[\ell(\kappa(p), y)]$  under the true labels equals the expected loss  $\mathbb{E}[\ell(\kappa(p), \tilde{y})]$  they would expect to suffer if  $P$  was indeed Bayes optimal.
- **No regrets:** The expected loss  $\mathbb{E}[\ell(\kappa(p), y)]$  under the true labels is indeed minimized by playing the best response  $\kappa^*$  over any other function  $\kappa : [0, 1] \rightarrow \mathcal{A}$ .

While our work focuses on calibration guarantees for decision making, in the the broader context, recent interest in calibration has been driven by the numerous surprising applications of calibration and its generalizations to algorithmic fairness [HKRR18], learning [GKR<sup>+</sup>22], complexity theory [CDV24], pseudorandomness [DKR<sup>+</sup>21], cryptography [HV25] and other areas of theoretical computer science.

Perfect calibration is an idealized notion; we cannot realistically expect it from predictors in the real world for computational and information-theoretic reasons. This has motivated the formulation of approximate notions of calibration [FV98, KF08, ZKS<sup>+</sup>21, BGHN23a, BGHN23b, BN24, HW24, OKK25, RSB<sup>+</sup>25]. To be useful, approximate notions of calibration should be computationally efficient, and yield some relaxed form of the guarantees above. We refer here to the efficiency of auditing for calibration error, where the goal is to estimate a calibration measure to within a prescribed additive error, from random samples of the form  $(p, y)$ . Following common practice in the literature (see e.g. [BGHN23a]), we henceforth drop  $x$  from our notation and consider the joint distribution  $(p, y)$ . While  $x$  influences how  $p = P(x)$  and  $y = y|x$  are jointly distributed, we do not mention it explicitly, to emphasize the fact that calibration measures (like loss functions) are typically independent of the feature space.

## 1.1 Approximate Calibration From Indistinguishability.

This is a view of calibration as a notion of indistinguishability between the real world and a simulation that the predictor  $P$  proposes. It is based on the outcome indistinguishability framework of [DKR<sup>+</sup>21] and developed in [GKSZ22, GH25]. The real world is modeled by the joint distribution  $J = (p, y)$  and the simulation by  $\tilde{J} = (p, \tilde{y})$ . Perfect calibration is equivalent to the two distributions being identical.

This naturally suggests relaxations that only require the distributions to be close, not identical, but also restrict the set of distinguishers to *fool*, in analogy with cryptography and pseudorandomness. This restriction turns out to be crucial for efficient auditing. This relaxation is captured by the definition of weight-restricted calibration [GKSZ22], where we consider a family of functions  $\mathcal{W} \subseteq \{[0, 1] \mapsto [-1, 1]\}$  and require that the distributions  $(p, y)$  and  $(p, \tilde{y})$  be indistinguishable to all predictors of the form  $\{f(p, y) = w(p)y\}_{w \in \mathcal{W}}$ .<sup>1</sup> Formally, for a distribution  $J = (p, y)$ , we define the  $\mathcal{W}$ -restricted calibration error as

$$\text{CE}_{\mathcal{W}}(J) = \max_{w \in \mathcal{W}} \mathbb{E}[w(p)(y - \tilde{y})] = \max_{w \in \mathcal{W}} \mathbb{E}[w(p)(y - p)].$$

The No Surprises property, which is referred to as decision OI in the literature [GHK<sup>+</sup>23], can be seen as a form of indistinguishability between these distributions.

A sequence of works in recent years [GKSZ22, BGHN23a, BGHN23b, BN24, OKK25, RSB<sup>+</sup>25] have painted a complete picture of weight families for which  $\mathcal{W}$ -restricted calibration is meaningful and tractable; we refer the reader to the recent survey [GH25]. We highlight a few results that are relevant to us:

1. **Complexity of auditing.** The tractability of estimating  $\text{CE}_{\mathcal{W}}$  is tightly captured by the learnability of the class  $\mathcal{W}$  [GHR24].
2. **Expected calibration Error.** The expected calibration error  $\text{ECE}(J)$  corresponds to  $\mathcal{W} = \mathcal{W}^*$  being all bounded functions. Since  $\mathcal{W}^*$  has unbounded VC dimension, ECE cannot be audited from finitely many samples [GHR24].
3. **Decision OI.** Decision OI is characterized by  $\text{CE}_{\text{Int}}$  where  $\text{Int}$  is the set of indicators of intervals in  $[0, 1]$  [OKK25].<sup>2</sup>
4. **Smooth calibration.** Considering the family of 1-Lipschitz weight functions gives smooth calibration [KF08], which is robust under small perturbations of the predictor, and captures an intuitive measure of calibration error called the distance to calibration [BGHN23a].

Since both intervals and Lipschitz functions are efficiently learnable in 1 dimension, the latter two notions can be audited efficiently (see also [HJTY24]).

## 1.2 Approximate Calibration From No Regret.

An alternate view of approximate calibration comes from relaxing the no regret property of perfect calibration. We will restrict our attention to (bounded) proper loss functions. Informally, for a loss function to be proper, a predictor with knowledge of the true outcome distribution should not be able to further decrease its loss by dishonestly predicting some alternate distribution. Formally, these

<sup>1</sup>All bounded distinguishers  $f(p, y)$  can be assumed to have the form  $w(p)y$ , see [GH25].

<sup>2</sup>This notion is called Proper calibration by [OKK25] and cutoff calibration by [RSB<sup>+</sup>25], we will refer to it as Interval-restricted calibration  $\text{CE}_{\text{Int}}$  in keeping with the weight-restricted calibration nomenclature.

are loss functions where the action space  $\mathcal{A}$  is the space of predictions  $[0, 1]$ , and when  $y \sim \text{Ber}(q)$ , the prediction  $p$  that minimizes the expected loss  $\mathbb{E}[\ell(p, y)]$  is  $p = q$ . Restricting to proper losses is without loss of generality since an arbitrary loss function  $\ell$  can be converted to a proper loss  $\ell'$  by composing it with the best response  $\ell' = \ell \circ \kappa_\ell^*$ .<sup>3</sup> We let  $\mathcal{L}^*$  denote the set of all bounded proper loss functions, and  $\mathcal{K}^* = \{[0, 1] \mapsto [0, 1]\}$  denote the family of all post-processings.

The starting point is the following characterization of calibration due to [FV98]:  $J = (p, y)$  is perfectly calibrated iff for every proper loss  $\ell \in \mathcal{L}^*$  and post-processing  $\kappa \in \mathcal{K}^*$ ,

$$\mathbb{E}[\ell(p, y)] \leq \mathbb{E}[\ell(\kappa(p), y)].$$

This means that in every miscalibrated predictor, this inequality is violated for some  $\ell, \kappa$  pair. The recent work of Hu and Wu [HW24], building on [KLST23], suggests using the magnitude of this violation as a measure of miscalibration, which they call the *calibration decision loss*. Formally, for a family of post-processings  $\mathcal{K} \subseteq \mathcal{K}^*$ ,<sup>4</sup> we define the *calibration decision loss relative to  $\mathcal{K}$*  as

$$\text{CDL}_{\mathcal{K}}(J) = \max_{\kappa \in \mathcal{K}} \mathbb{E}[\ell(p, y) - \ell(\kappa(p), y)].$$

This leads to a natural family of calibration measures parametrized by  $\mathcal{K}$ , which measures how much regret (excess loss) a decision maker could possibly suffer for a lack of calibration, relative to a baseline set of post-processings  $\mathcal{K}$ . In this work we propose studying the complexity of  $\text{CDL}_{\mathcal{K}}$  for general families  $\mathcal{K}$  of post-processing functions.

Motivated by the online prediction setting, [HW24] were primarily interested in the case  $\mathcal{K} = \mathcal{K}^*$ , where they show a tight connection with the ECE:

$$\text{ECE}(J)^2 \lesssim \text{CDL}_{\mathcal{K}^*}(J) \lesssim \text{ECE}(J).$$

In the online setting, they showed surprisingly that CDL admits much lower regret rates than ECE. However, for the auditing problem in the offline setting which is our main focus, this tight connection to ECE means that  $\text{CDL}_{\mathcal{K}^*}$  cannot be audited from finitely many samples. One might hope to circumvent this intractability by suitably restricting the family of post-processings  $\mathcal{K}$ , in analogy to the weight-restricted calibration setting. The family of monotone post-processings  $\mathcal{M}_+$  was considered in [RSB<sup>+</sup>25], who bound the calibration decision loss relative to  $\mathcal{M}_+$  by  $\text{CE}_{\text{Int}}$ , where  $\text{Int}$  is the family of indicators of intervals, showing that  $\text{CDL}_{\mathcal{M}_+}(J) \leq 2\text{CE}_{\text{Int}}(J)$ .<sup>5</sup>

In addition to being a natural question from a theoretical viewpoint, understanding  $\text{CDL}_{\mathcal{K}}$  for various classes  $\mathcal{K}$  is relevant to how calibration errors are measured and remediated in practice [GPSW17]. Popular methods for recalibration such as Isotonic Regression [ZE01, ZE02] and the Pool Adjacent Violators (PAV) algorithm [ABE<sup>+</sup>55] and Platt scaling [Pla00], try to find the best post-processing from a family  $\mathcal{K}$  (see [NC05] for more details about these methods). Isotonic regression (which is the problem solved by PAV) considers monotone post-processings, whereas Platt scaling considers a parametrized subclass of monotone functions consisting of sigmoids. Implicitly, such methods consider the predictor calibrated if  $\text{CDL}_{\mathcal{K}}$  is small. Other families of post-processings like recalibration based on Uniform-mass binning [ZE01, GR21, SSH23] and vector/matrix scaling have been proposed [GPSW17]. Methods that find the best post-processing from a small class have also shown to be effective in practice (see e.g. [BHJB25]). However, there was a lack of theory to guide the choice of  $\mathcal{K}$ .

<sup>3</sup>In other words, a predictor that knows that an agent will best-respond to their predictions can view the agent as optimizing a proper loss.

<sup>4</sup>We assume  $\mathcal{K}$  contains the identity function, which ensure the positivity of  $\text{CDL}_{\mathcal{K}}$ .

<sup>5</sup>While not stated in terms of CDL, Proposition 3.2 in their paper is equivalent to the above inequality.

Other results in the literature on decision making from calibration correspond to studying CDL for restrictions of post-processings, loss functions or both. For instance, [BGHN23b] consider Lipschitz post-processings, and show that the maximum improvement to the expected squared loss  $\ell_2(p, y) = (y - p)^2$  from such post-processings is quadratically related to the smooth calibration error. But in contrast to weight-restricted calibration, there wasn't a comprehensive understanding of when calibration decision loss relative to post-processings in  $\mathcal{K}$  is a tractable measure.

**Our work:** In this work, we seek to understand the families of post-processings  $\mathcal{K}$  for which  $\text{CDL}_{\mathcal{K}}$  gives a tractable calibration measure that has strong guarantees for downstream decision makers. This raises several natural questions:

1. **Complexity characterization.** For what families of post-processings  $\mathcal{K}$  is estimating  $\text{CDL}_{\mathcal{K}}$  tractable in terms of sample and computational complexity? Is there a complexity measure for  $\mathcal{K}$  that governs tractability?
2. **Specific post-processings.** Is CDL estimation tractable for monotone post-processings and Lipschitz post-processings?<sup>6</sup> What is the most-expressive family  $\mathcal{K}$  for which we can estimate  $\text{CDL}_{\mathcal{K}}$  efficiently?
3. **Efficient post-processing.** If a predictor has large calibration decision loss relative to  $\mathcal{K}$ , how should we post-process it?
4. **Relation to weight-restricted calibration.** How does calibration decision loss relative to  $\mathcal{K}$  relate to weight-restricted calibration? Does small calibration error in one sense imply small error in the other?

The main contribution of this paper is to develop a comprehensive theory of when calibration decision loss relative to  $\mathcal{K}$  is tractable. We use this theory to answer the questions raised above. Some highlights from our results include:

- We show that allowing all Lipschitz post-processings results in a CDL notion that is information-theoretically hard to estimate, requiring unbounded sample size. This is in stark contrast to weighted calibration, where allowing all Lipschitz weight functions yields smooth calibration, which is not only tractable but captures distance to calibration and anchors the notion of consistent calibration measures as defined by [BGHN23a].
- We show that restricting post-processings to monotone functions and their generalizations yield tractable notions of CDL. Our theoretical results justify the central role such families play in practice [Pla00, ABE<sup>+</sup>55, ZE02].
- We establish tight connections between calibration decision loss and weight restricted calibration for any valid post-processing class  $\mathcal{K}$ , unifying and generalizing the previous results of [KLST23, HW24, RSB<sup>+</sup>25].
- We introduce new algorithmic techniques from the omniprediction literature to post-processing. We give rigorous new guarantees for some commonly used recalibration procedures in the machine learning literature like Isotonic regression/Pool Adjacent Violators [ABE<sup>+</sup>55, ZE02] and Uniform mass binning [ZE01, GR21, SSH23].

---

<sup>6</sup>We invite the reader to make a guess about these two specific families before reading further.

## 2 Overview of Our Results

In this section, we present the key definitions and an overview of our main results, and highlight some of the key ideas behind them, without getting into the technical details.

### 2.1 Definitions

#### Proper loss functions.

**Definition 2.1** (Proper Losses). A loss  $\ell : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}$  is *proper* if for all  $p, q \in [0, 1]$ ,

$$\mathbb{E}_{y \sim \text{Ber}(q)} [\ell(p, y) - \ell(q, y)] \geq 0.$$

We let  $\mathcal{L}^*$  be the class of proper losses such that  $|\ell(p, 1) - \ell(p, 0)| \leq 1$  for all  $p \in [0, 1]$ .<sup>7</sup>

Of particular importance are the *V-shaped* losses studied in [HSLW23, KLST23], which are all functions

$$\ell_v(p, y) = -\text{sign}(p - v)(y - v),$$

where  $v \in [0, 1]$ . These are proper losses, and moreover they form a basis for  $\mathcal{L}^*$ ; a precise formulation which builds on [KLST23] is given in Lemma 3.5.

**Post-processing functions.** We say that a class of post-processings  $\kappa : [0, 1] \rightarrow [0, 1]$  is *valid* if it contains the identity function and satisfies a certain *translation invariance* under shifts of either axis (the precise definition is in Definition 3.9). Most natural classes of post-processings we know that have been considered previously, including  $\mathcal{K}^*$ , Lipschitz, monotone and bounded degree polynomial post-processings, are valid classes. For a valid post-processing class,  $\mathcal{K}$ , let

$$\text{thr}(\mathcal{K}) = \left\{ p \mapsto \text{sign}_+ \left( \kappa(p) - \frac{1}{2} \right) \mid \kappa \in \mathcal{K} \right\}.$$

While the choice of  $1/2$  might seem arbitrary, the translation invariance of  $\mathcal{K}$  implies that any constant in  $(0, 1)$  will do.

We will pay special attention to the class of *monotonic* post-processing functions, as well as a natural generalization of this class. To define the class, recall that an *interval*  $I \subseteq [0, 1]$  with no further specification may be open, closed, half-open, or a singleton.

**Definition 2.2** (Generalized Monotonicity). Given an integer  $r \in \mathbb{N}$  and a function  $\kappa : [0, 1] \rightarrow [0, 1]$ , we say  $\kappa \in \mathcal{M}_r$  if for all values  $v \in \mathbb{R}$ , the *v-superlevel set* of  $\kappa$ ,

$$\kappa^{-1}([v, 1]) = \{p \in [0, 1] \mid \kappa(p) \geq v\},$$

can be expressed as the union of at most  $r$  disjoint intervals  $I_1, \dots, I_r \subseteq [0, 1]$ . Then,  $\mathcal{M}_r$  is a valid post-processing class. In addition, let  $\mathcal{M}_+$  and  $\mathcal{M}_-$  denote the sets of monotonically nondecreasing and nonincreasing functions  $\kappa : [0, 1] \rightarrow [0, 1]$ , respectively.

We call  $\mathcal{M}_r$  a class of “generalized” monotonic functions because the monotonic functions  $\mathcal{M}_+$  and  $\mathcal{M}_-$  are both subsets of  $\mathcal{M}_1$ , and because  $\mathcal{M}_r \subseteq \mathcal{M}_s$  for all  $r \leq s$ . We also note that  $\mathcal{M}_r$  and  $\mathcal{M}_+$  are valid post-processing classes, but  $\mathcal{M}_-$  is not: Although  $\mathcal{M}_-$  is translation invariant, it does not contain the identity function. Generalized monotone functions are the broadest class of weight functions that admit efficient algorithms for CDL by our results. In Section A we discuss why they constitute a natural class of post-processings to consider, both from a theoretical and practical viewpoint. In Figure 1 we give an example of a function which is (very) non-monotone, but is a 3-generalized monotone function.

---

<sup>7</sup>Note that our definition of  $\mathcal{L}^*$  includes all proper loss functions with range  $[0, 1]$ . The exact range  $([-1, 1]$  versus  $[0, 1])$  is not crucial, it will only change the definition of CDL by a factor of 2.

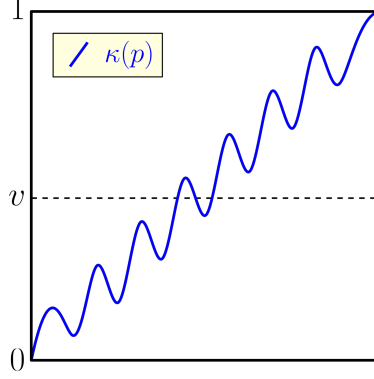


Figure 1: Function  $\kappa : [0, 1] \rightarrow [0, 1]$  that crosses every threshold  $v \in [0, 1]$  of its range at most 3 times. The monotonicity of the function  $\kappa(p)$  changes 14 times as  $p$  grows from 0 to 1. Although  $\kappa$  is non-monotone, we have  $\kappa \in \mathcal{M}_r$  for  $r = 3$ .

**Calibration Decision Loss.** Next we define *calibration decision loss* (CDL), introduced by [HW24]. We study a version of CDL that is parameterized by a *valid post-processing class*  $\mathcal{K}$ .

**Definition 2.3** (Calibration Decision Loss relative to  $\mathcal{K}$ ). Given valid post-processing class  $\mathcal{K}$  and a joint distribution  $J$  over pairs  $(p, y) \in [0, 1] \times \{0, 1\}$ , the *calibration decision loss* is

$$\text{CDL}_{\mathcal{K}}(J) = \sup_{\ell, \kappa} \mathbb{E}[\ell(p, y) - \ell(\kappa(p), y)],$$

where the supremum is taken over all  $\ell \in \mathcal{L}^*$  and  $\kappa \in \mathcal{K}$ . Given a subset  $\mathcal{L} \subset \mathcal{L}^*$  we define the  $(\mathcal{L}, \mathcal{K})$ -calibration decision loss as

$$\text{CDL}_{\mathcal{L}, \mathcal{K}}(J) = \sup_{\ell \in \mathcal{L}, \kappa \in \mathcal{K}} \mathbb{E}[\ell(p, y) - \ell(\kappa(p), y)].$$

Given a particular  $\ell \in \mathcal{L}^*$ , we also define a loss-specific version of CDL, called the *calibration fixed decision loss*, as

$$\text{CDL}_{\ell, \mathcal{K}}(J) = \sup_{\kappa} \mathbb{E}[\ell(p, y) - \ell(\kappa(p), y)].$$

For any valid  $\mathcal{K}$ ,  $\text{CDL}_{\mathcal{K}}$  is always non-negative since the identity function is in  $\mathcal{K}$ . When  $\mathcal{K} = \mathcal{K}^*$  is the family of *all* functions  $\kappa : [0, 1] \rightarrow [0, 1]$ , then our notion of CDL coincides with that of [HW24].

In the remainder of this section, we describe the three main tasks related to calibration decision loss that we study in our work: efficient testing/auditing, relationship to weight-restricted calibration and efficient omniprediction, and our results for each of them.

## 2.2 Testing and Auditing

The first problem we study is the characterization of post-processing classes  $\mathcal{K}$  that allow for efficient *testing* of  $\text{CDL}_{\mathcal{K}}$ , which we define as follows.

**Definition 2.4** (Testing CDL). We say that an algorithm  $\mathcal{A}$  is an  $(\alpha, \beta)$ -tester for  $\text{CDL}_{\mathcal{K}}$ , if the algorithm, upon receiving a large enough set of i.i.d. examples from some unknown distribution  $J$ , outputs either **Accept** or **Reject**, and satisfies the following properties with probability at least  $2/3$ .

1. If  $\text{CDL}_{\mathcal{K}}(J) > \alpha$ , then  $\mathcal{A}$  outputs **Reject**,



2. If  $\text{CDL}_{\mathcal{K}}(J) \leq \beta$ , then  $\mathcal{A}$  outputs **Accept**.

As is typical, the probability here is taken over the sampling process, and the internal randomness of the tester. The tester's output is unconstrained if  $\text{CDL}_{\mathcal{K}}(J) \in (\beta, \alpha)$ . We also consider a relaxation of the testing problem, called *auditing*, which we define in terms of the *expected calibration error*  $\text{ECE}(J) = \mathbb{E}|\mathbb{E}[y|p] - p|$ . Recall that Hu and Wu [HW24] showed that  $\text{CDL}_{\mathcal{K}}(J) \leq 2 \cdot \text{ECE}(J)$ . The following definition weakens the second requirement to accepting predictors with small ECE.

**Definition 2.5** (Auditing for CDL). We say that an algorithm  $\mathcal{A}$  is an  $(\alpha, \beta)$ -auditor for  $\text{CDL}_{\mathcal{K}}$ , if the algorithm, upon receiving a large enough set of i.i.d. examples from some unknown distribution  $J$ , outputs either **Accept** or **Reject**, and satisfies the following property, with probability at least  $2/3$ .

1. If  $\text{CDL}_{\mathcal{K}}(J) > \alpha$ , then  $\mathcal{A}$  outputs **Reject**,
2. If  $\text{ECE}(J) \leq \beta/2$ , then  $\mathcal{A}$  outputs **Accept**.

**Sample complexity of testing and auditing.** Our first result shows that the VC dimension of  $\text{thr}(\mathcal{K})$  governs the sample efficiency of calibration testing and auditing for  $\text{CDL}_{\mathcal{K}}$ .

**Theorem 2.6** (Sample Complexity Bounds). *Let  $\mathcal{K}$  be a valid post-processing class, and let  $d = \text{VCdim}(\text{thr}(\mathcal{K}))$ . Then,*

1. *For any  $\alpha, \varepsilon \in (0, 1)$ , there is an  $(\alpha, \alpha - \varepsilon)$ -tester for  $\text{CDL}_{\mathcal{K}}$  with sample complexity  $\tilde{O}(d/\varepsilon^2)$ .*
2. *Any  $(1/8, 0)$ -auditor for  $\text{CDL}_{\mathcal{K}}$  requires  $\Omega(\sqrt{d})$  samples.*

Some implications of this result:

- Auditing is easier than testing, since any  $(\alpha, \beta)$ -tester for  $\text{CDL}_{\mathcal{K}}$  is also an  $(\alpha, \beta)$ -auditor. Hence our lower bound for auditing also applies to testers. The case  $\beta = 0$  corresponds to perfect calibration, so our lower bound holds even for algorithms that must only accept perfectly calibrated predictors, and reject predictors with large  $\text{CDL}_{\mathcal{K}}$ .
- A key corollary is that auditing  $\text{CDL}_{\text{Lip}}$  is not possible with finitely many samples. This follows from the observation that  $\text{thr}(\text{Lip})$  has unbounded VC dimension (see [Corollary 4.4](#)).
- The lower bound makes use of V-shaped losses. These losses are not strongly convex unlike common proper losses used in practice, and one may wonder if the lower bound can be circumvented by assuming strong convexity. The answer is No: we prove that essentially the same lower bound carries over even for strongly convex losses (see [Corollary 4.7](#)).
- V-shaped losses are discontinuous in the prediction value  $p$ . If we restrict our attention to loss functions that are Lipschitz continuous as a function of  $p$ , and only consider the CDL for such losses and Lipschitz post-processings, then we show that this measure is quadratically related to the smooth calibration error, hence it is efficiently auditable (see [Theorem 4.9](#)).

**Computationally efficiency from learnability.** We show that efficient algorithms for testing and auditing  $\text{CDL}_{\mathcal{K}}$  can be derived from efficient learning algorithms for  $\text{thr}(\mathcal{K})$ .

We will use the standard primitive of agnostic learning, first introduced by [KSS94]. It is known to be equivalent to seemingly weaker primitives like weak agnostic learning [Fel09, KK09].

**Definition 2.7** (Agnostic Learning). Let  $\mathcal{C} \subseteq \{[0, 1] \rightarrow \{\pm 1\}\}$ . We say that an algorithm  $\mathcal{A}$  is an  $\varepsilon$ -agnostic learner for  $\mathcal{C}$  if, upon receiving a large enough set of i.i.d. examples from some unknown distribution  $D$  over  $[0, 1] \times \{\pm 1\}$ ,  $\mathcal{A}$  outputs some hypothesis  $h : [0, 1] \rightarrow \{\pm 1\}$  such that, with probability at least 0.9 over the samples and the internal randomness of  $\mathcal{A}$ , we have:

$$\mathbb{P}_{(p,z) \sim D}[h(p) \neq z] \leq \min_{f \in \mathcal{C}} \mathbb{P}_{(p,z) \sim D}[f(p) \neq z] + \varepsilon$$

Moreover, if  $h \in \mathcal{C}$  we say that  $\mathcal{A}$  is proper.

Our main algorithmic result is a reduction from  $\text{CDL}_{\mathcal{K}}$  testing to proper agnostic learning for  $\text{thr}(\mathcal{K})$ . For  $\text{CDL}_{\mathcal{K}}$  auditing, improper agnostic learning suffices.

**Theorem 2.8** (Testing and Auditing from Agnostic Learning, [Theorem 6.1](#)). *Let  $\mathcal{K}$  be a valid post-processing class. Let  $\text{AL}$  be an  $\varepsilon$ -agnostic learner for  $\text{thr}(\mathcal{K})$ . Then, for any  $\alpha \in (0, 1)$ , there is an  $(\alpha, \alpha - 3\varepsilon)$ -auditor for  $\text{CDL}_{\mathcal{K}}$  that makes  $\tilde{O}(1/\varepsilon)$  calls to  $\text{AL}$ . Moreover, if  $\text{AL}$  is proper, then there is an  $(\alpha, \alpha - 3\varepsilon)$ -tester for  $\text{CDL}_{\mathcal{K}}$  of similar complexity.*

This result lets us translate efficient agnostic learning algorithms for classes of Boolean functions on  $[0, 1]$  to efficient algorithms for  $\text{CDL}_{\mathcal{K}}$  estimation for valid post-processing classes.

- For the class  $\mathcal{M}_+$  of monotone post-processings,  $\text{thr}(\mathcal{M}_+)$  is the family of intervals  $[a, 1]$ , and it is a classic result that intervals are efficiently agnostically learnable. The efficient tester that results from [Theorem 2.8](#) strengthens and extends the result of [\[RSB<sup>+</sup>25\]](#) who showed that  $\text{CDL}_{\mathcal{M}_+}$  can be bounded by ensuring low weight-restricted calibration error for intervals.
- More generally, for  $\kappa \in \mathcal{M}_r$ , the sets  $\kappa(p) \geq v$  can be expressed as the union of at most  $r$  disjoint intervals. We show that this class admits an efficient proper agnostic learner. As a corollary, we obtain an efficient tester for  $\text{CDL}_{\mathcal{M}_r}$ .

## 2.3 Calibration Decision Loss and Weight-Restricted Calibration

The next problem we study is the relation between  $\text{CDL}_{\mathcal{K}}$  and *weight-restricted* calibration measures, defined below. The standard definition [\[GKSZ22, GH25\]](#) either takes the absolute value of the expectation or assumes that  $\mathcal{W}$  is closed under negation, we intentionally will do neither.

**Definition 2.9** (Weight-Restricted Calibration Error). Given a distribution  $J$  over  $(p, y) \in [0, 1] \times \{0, 1\}$  and a class of weight functions  $\mathcal{W} \subseteq \{[0, 1] \rightarrow [-1, +1]\}$ , the  $\mathcal{W}$ -restricted calibration error is

$$\text{CE}_{\mathcal{W}}(J) = \sup_{w \in \mathcal{W}} \mathbb{E}[w(p)(y - p)].$$

In the case that  $\mathcal{W}$  is closed under negations, we have  $\text{CE}_{\mathcal{W}}(J) \geq 0$ .

Several works [\[KLST23, HW24, BGHN23b, RSB<sup>+</sup>25\]](#) have proved results that can be viewed as instances of this general question for specific choice of loss families  $\mathcal{L}$  and post-processings  $\mathcal{K}$ . Such characterizations are valuable because weight-restricted calibration error measures have been well studied in the literature, and we understand the relation between various measures fairly well [\[BGHN23a, GH25\]](#). Moreover, we would like notions of approximate calibration to simultaneously give small calibration decision loss and strong indistinguishability, making it natural to ask to what extent one implies the other.

We significantly extend and complete this line of work with a general characterization that holds for all valid classes  $\mathcal{K}$ .

**Theorem 2.10.** *Given a valid post-processing class  $\mathcal{K}$ , let*

$$\text{thr}'(\mathcal{K}) = \text{thr}(\mathcal{K}) \cup \left\{ p \mapsto -\text{sign}_+(p - v) \mid v \in \mathbb{R} \right\}.$$

*Then*

$$\text{CE}_{\text{thr}'(\mathcal{K})}(J)^2 \lesssim \text{CDL}_{\mathcal{K}}(J) \lesssim \text{CE}_{\text{thr}'(\mathcal{K})}(J).$$

While the class  $\text{thr}(\mathcal{K})$  contains only the *upper* thresholds of functions in  $\mathcal{K}$ , the class  $\text{thr}'(\mathcal{K})$  also includes all *lower* thresholds of the identity function. It also contains upper thresholds of the identity function by the inclusion of  $\text{thr}(\mathcal{K})$ , since  $\mathcal{K}$  is translation invariant and contains the identity. Since our definition of weight-restricted calibration error did not involve absolute values, the decision for  $\text{thr}'(\mathcal{K})$  to not include lower thresholds of all functions in  $\mathcal{K}$  is deliberate and consequential.

This change in the definition of the appropriate concept class for characterizing testability and for the characterization in terms of weighted calibration error may at first seem counterintuitive. This arises because CDL is fundamentally an asymmetric notion (meaning we do not assume  $\kappa \in \mathcal{K}$  implies  $1 - \kappa \in \mathcal{K}$ ); e.g. monotone increasing post-processings are significantly more natural than monotone decreasing ones. Our definition of  $\text{thr}'(\mathcal{K})$  precisely captures how much we need to enhance  $\text{thr}(\mathcal{K})$  to be able to relate it to weighted calibration. Our results on testing, auditing and omniprediction are characterized by naturally symmetric notions such as VC Dimension, and agnostic learning, and would hold equally well if we used  $\text{thr}'(\mathcal{K})$  instead of  $\text{thr}(\mathcal{K})$ , or even if we added all lower thresholds of functions in  $\mathcal{K}$ ; this would not change the VC dimension or agnostic learnability. But [Theorem 2.10](#) would no longer hold with  $\text{thr}(\mathcal{K})$  in place of  $\text{thr}'(\mathcal{K})$ , as explained in the proof sketch below.

This result generalizes and significantly extends results that were previously known in the literature. For instance, when  $\mathcal{K} = \mathcal{K}^*$ , then  $\text{thr}(\mathcal{K})$  consists of all functions  $\mathcal{W}^* = \{[0, 1] \mapsto [-1, 1]\}$  and the corresponding weight-restricted calibration notion is just ECE. Thus in this case, we recover the result of [\[HW24, KLST23\]](#) which shows that

$$\text{ECE}(J)^2 \lesssim \text{CDL}_{\mathcal{K}^*}(J) \lesssim \text{ECE}(J).$$

When  $\mathcal{K} = \mathcal{M}_+$  is all monotone increasing functions, then  $\text{thr}(\mathcal{M}_+)$  contains all indicators of intervals  $[a, 1]$  for  $a \in [0, 1]$ . Let  $\text{Int}$  denote the collection of these indicators, along with their complements. In this case, our result shows that

$$\text{CE}_{\text{Int}}(J)^2 \lesssim \text{CDL}_{\mathcal{M}_+}(J) \lesssim \text{CE}_{\text{Int}}(J).$$

The upper bound was shown in the work of [\[RSB<sup>+</sup>25\]](#), while the lower bound is new. For the class of generalized monotone functions  $\mathcal{M}_r$ , our theorem shows that  $\text{CDL}_{\mathcal{M}_r}$  is quadratically related to a generalized version of  $\text{CDL}_{\text{Int}}$ , in which single intervals are replaced by unions of at most  $r$  intervals.

**Proof Sketch.** [Theorem 2.10](#) is a highly general result, which transforms a bound on the  $\text{thr}'(\mathcal{K})$ -restricted calibration error into a bound on  $\text{CDL}_{\mathcal{K}}$ , and vice versa. In the language of outcome indistinguishability, the key insight underlying the proof is that the two types of functions in the weight class  $\text{thr}'(\mathcal{K})$ , namely upper thresholds of  $\mathcal{K}$  and lower thresholds of the identity, are precisely what we need to move, respectively, *to* and *from* the world of simulated outcomes. This perspective builds on the *loss OI* framework of [\[GHK<sup>+</sup>23\]](#), particularly their study of *decision OI*.

In slightly more detail, we first consider a weight function obtained by composing a post-processing function  $\kappa(p)$  with a proper loss function's *negated discrete derivative*  $-\partial\ell(p) = \ell(p, 0) - \ell(p, 1)$ . In the worst case—that is, the case of a V-shaped loss function—this composite function corresponds

exactly to some *upper threshold* of  $\kappa$ . Calibration with respect to this composite weight function ensures that moving from real outcomes  $y$  to simulated outcomes  $\tilde{y}$  can only *improve* the loss attained by  $\kappa(p)$ , up to an  $\varepsilon$  slack factor:

$$\mathbb{E}[\ell(\kappa(p), y)] \geq \mathbb{E}[\ell(\kappa(p), \tilde{y})] - \varepsilon.$$

Next, we observe that in the simulated world, where  $\tilde{y} \sim \text{Ber}(p)$ , the predictor  $p$  is Bayes optimal by definition. In particular, it outperforms  $\kappa(p)$ :

$$\mathbb{E}[\ell(\kappa(p), \tilde{y})] \geq \mathbb{E}[\ell(p, \tilde{y})].$$

Finally, we will show that weight functions corresponding to *lower thresholds* of the identity allow us to move back to the world of real outcomes, again up to a small slack factor:

$$\mathbb{E}[\ell(p, \tilde{y})] \geq \mathbb{E}[\ell(p, y)] - \varepsilon.$$

Combining this chain of inequalities, we deduce that  $p$  outperforms  $\kappa(p)$  *in the real world, as well*. Phrased differently,  $\text{thr}'(\mathcal{K})$ -restricted calibration ensures that  $\text{CDL}_{\mathcal{K}}$  is small.

The proof of the converse implication—that small  $\text{CDL}_{\mathcal{K}}$  implies small  $\text{thr}'(\mathcal{K})$ -restricted calibration error (up to a quadratic gap)—relies on similar ideas, but is much subtler. For simplicity, consider the most fundamental post-processing class for which the result was not previously known: monotonically increasing functions  $\mathcal{K} = \mathcal{M}_+$ . In this case, we are given that  $\text{CDL}_{\mathcal{M}_+} \leq \varepsilon$  and are tasked with proving that  $\text{CE}_{\text{Int}} \lesssim \sqrt{\varepsilon}$ . To do so, we consider an arbitrary interval  $I \subseteq [0, 1]$  and break it into  $m = O(1/\sqrt{\varepsilon})$  subintervals  $I_1, \dots, I_m$  with roughly equal probability mass under the distribution of predictions. We then show that the calibration error restricted to a particular subinterval  $I_j = [a, b]$  can be bounded by the product of its length and mass, plus an  $\varepsilon$  slack term. Since each subinterval has mass  $\approx 1/m$  and their lengths sum to at most 1, our total bound becomes roughly  $1/m + m\varepsilon = O(\sqrt{\varepsilon})$ . We give our full proof of [Theorem 2.10](#) in [Section 5](#).

We conclude by observing that some assumption on  $\mathcal{K}$  like validity is necessary for the characterization to hold: the characterization does not hold for the class  $\mathcal{M}_-$  of non-increasing post-processings. This class is translation invariant, but does not contain the identity, so it is not valid by our definition.

## 2.4 Post-Processing and Omniprediction

Here we ask the question: if the predictor  $P$  suffers large calibration decision loss relative to  $\mathcal{K}$ , how should we remedy it? We would ideally like to post-process it in an efficient manner that gives guarantees for every loss in  $\mathcal{L}^*$ , competitive to baselines from the set  $\mathcal{K}$ . But the issue is that there might be several losses in  $\ell \in \mathcal{L}^*$  that witness large calibration decision loss, each with its own post-processing  $\kappa = \kappa^\ell$ , which might not be good for a different loss.

To circumvent this, we could allow post-processings that need not themselves lie in  $\mathcal{K}$ . For instance, if we could take  $\kappa(p) = \mathbb{E}[y|p]$  to be the perfect recalibration, then we would have a guarantee for all losses. However, this function will be inefficient to compute (it is as hard as estimating ECE). Under what conditions can we efficiently find a post-processing that gives guarantees for all losses?

We formulate this as a problem of efficient *omniprediction* [\[GKR<sup>+</sup>22, GHK<sup>+</sup>23\]](#). The notion of omniprediction originating in supervised learning [\[GKR<sup>+</sup>22\]](#) asks for a predictor that is simultaneously competitive with the best hypothesis from a class  $\mathcal{C}$  of hypotheses, for any loss from a family  $\mathcal{L}$ . The power of this notion comes from the fact that the best hypothesis in  $\mathcal{C}$  can depend

on the loss  $\ell \in \mathcal{L}$ , whereas our predictor is oblivious to  $\ell$ . In our context, the goal is to learn a post-processing function  $\hat{\kappa}$  that outperforms any other in  $\mathcal{K}$  with respect to *all* possible decision tasks. This is a departure from its standard definition, where the baseline is a hypothesis class  $\mathcal{C}$  of functions on the feature space  $\mathcal{X}$ , and is similar to its formulation in the recent work of [HWY25], that shows omniprediction guarantees for smooth calibration.

**Definition 2.11** (Omniprediction). We say that a function  $\hat{\kappa} : [0, 1] \rightarrow [0, 1]$  is an  $(\varepsilon, \mathcal{K})$ -omnipredictor for some distribution  $J$  over pairs  $(p, y)$  if for all  $\ell \in \mathcal{L}^*$  and  $\kappa \in \mathcal{K}$ ,

$$\mathbb{E}[\ell(\hat{\kappa}(p), y)] \leq \mathbb{E}[\ell(\kappa(p), y)] + \varepsilon.$$

We say that  $\mathcal{A}$  learns an  $(\varepsilon, \mathcal{K})$ -omnipredictor with probability  $1 - \delta$  if, upon receiving a large enough set of i.i.d. samples from some unknown distribution  $J$ , the algorithm  $\mathcal{A}$  outputs an  $(\varepsilon, \mathcal{K})$ -omnipredictor for  $J$  with probability at least  $1 - \delta$ .

In this section, we will suppress the dependence on  $\varepsilon, \delta$  and say that an algorithm learns a  $\mathcal{K}$ -omnipredictor if it returns an  $(\varepsilon, \mathcal{K})$ -omnipredictor with high probability.<sup>8</sup> We show a range of omniprediction guarantees for various classes  $\mathcal{K}$ , either by using techniques from the omniprediction literature [GHK<sup>+</sup>23], or by a new analysis of well-known recalibration procedures that have been proposed in the machine learning literature [ABE<sup>+</sup>55, ZE01, GR21, SSH23]. We start with the most general result.

**Omniprediction from agnostic learning.** For all valid post-processing classes  $\mathcal{K}$ , we prove that an omnipredictor can be efficiently learnt, under the assumption that  $\text{thr}(\mathcal{K})$  is agnostically learnable. Thus the same assumption we require for efficient auditing of  $\text{CDL}_{\mathcal{K}}$  is in fact sufficient to ensure omniprediction.

**Theorem 2.12** (Omniprediction from Agnostic Learning, Informal Version of Theorem 7.1). *For every valid post-processing class  $\mathcal{K}$ , there is an efficient reduction from learning a  $\mathcal{K}$ -omnipredictor to agnostic learning for  $\text{thr}(\mathcal{K})$ .*

We follow the loss OI framework of [GHK<sup>+</sup>23] which is an indistinguishability-based approach to omniprediction. The key difference between their setting and ours is that they compete against a baseline of hypotheses  $\mathcal{C} = \{c : \mathcal{X} \rightarrow \{+1, -1\}\}$  where  $\mathcal{X}$  denotes the feature space. Whereas in our calibration setting, we do not have a feature space  $\mathcal{X}$ , we compete against a baseline of post-processing functions  $\kappa(p)$ . Nevertheless, we show how one can adapt their techniques to the setting of calibration to learn omnipredictors efficiently.

**Pool Adjacent Violators is an omnipredictor.** The Pool Adjacent Violators algorithm [ABE<sup>+</sup>55] solves the problem of isotonic regression: given samples from  $J = (p, y)$ , it finds a monotone post-processing of a predictor  $p$  that minimizes the square loss among all monotone post-processings. Given a sample of  $\{(y_i, p_i)\}$  pairs, it starts from the Bayes optimal predictor on the sample  $\kappa(p) = \mathbb{E}[y|p]$ , and pools/merges any adjacent pair that violates monotonicity, till it reaches a monotone predictor. We present the algorithm formally in Algorithm 1.

Various works have observed that it actually gives guarantees for broader classes of loss functions (including convex proper losses); see [ZE02, BdP13] and references therein. We will show that it is an omnipredictor for all of  $\mathcal{L}^*$  relative to monotone post-processings. This general statement is

<sup>8</sup>The results will typically involve a reduction to some other algorithm, whose parameters will be suitably chosen functions of  $\varepsilon, \delta$ . The formal theorem statements make this dependence explicit.

new to our knowledge. For instance [BdP13] shows the result for a subset of scoring rules that they call regular proper scoring rules, defined in terms of certain integrals. This does not capture all proper losses (indeed it is unclear to us what subset they capture), and they use considerably more complex arguments. We present a simple argument that relies on the *closer is better* property of proper losses (see Lemma 3.2) which states for a proper loss, moving the prediction  $p$  closer to the Bayes optimal  $\mathbb{E}[y|p]$  can only help. Formally, we establish the following:

**Theorem 2.13** (Omniprediction through PAV, Informal Version of Theorem 7.5). *Pool Adjacent Violators (PAV) with sufficiently many samples learns a  $\mathcal{M}_+$  omnipredictor.*

This shows that the class of monotone post-processings admits a *proper omnipredictor*: for every distribution  $J = (p, y)$ , there is a single post-processing  $\kappa^* \in \mathcal{M}_+$  with the guarantee that for every proper loss  $\ell \in \mathcal{L}^*$ , and  $\kappa \in \mathcal{M}_+$ ,

$$\mathbb{E}[\ell(\kappa^*(p), y)] \leq \mathbb{E}[\ell(\kappa(p), y)].$$

Other than  $\mathcal{K}^*$ , this is the only natural class of post-processings we know that has this property.

**Omniprediction through Bucketing and Recalibration.** We next analyze bucketed recalibration through uniform-mass binning. Binning is a long-established technique for measuring calibration [Mil62, San63]. The method of uniform-mass binning was introduced by [ZE01] as the first binning-based approach not only for measuring calibration, but also for obtaining a calibrated predictor. Rather than choosing equal-width bins, we choose bin boundaries as quantiles, so that every bin has roughly the same mass. It remains empirically competitive to this day [NCH15, GPSW17, RCSM22], and its calibration properties have been theoretically studied as well [GR21, SSH23].

Informally, uniform-mass binning tries to divide  $[0, 1]$  into bins that each have probability  $\varepsilon$  of containing the prediction  $p$ . If  $p$  has a continuous distribution, we can do this by taking the  $\varepsilon$ -quantiles as bin boundaries. For general distributions, we might need to allow *singleton* buckets consisting of a single point to account for point-masses that might be larger than  $\varepsilon$ . This gives a partition of  $[0, 1]$  into  $O(1/\varepsilon)$  buckets that are either singletons, or have probability bounded by  $\varepsilon$ .

We show that uniform-mass binning followed by recalibration yields omniprediction with respect to the class of generalized monotone post-processings.

**Theorem 2.14** (Omniprediction from Uniform-mass binning, Informal Version of Theorem 7.9). *For all  $r \geq 1$ , Uniform-mass binning (with  $O(r^2)$  bins) and recalibration (Algorithm 2) learns an  $\mathcal{M}_r$ -omnipredictor.*

The proof proceeds by comparing the bucket-wise recalibration  $\hat{\kappa}$  to any hypothesis in  $\kappa \in \mathcal{M}_r$ , for a specific V-shaped loss  $\ell_v$ . We rely on two observations:

1. If  $\text{sign}(\kappa(p) - v)$  is constant for a bucket  $I_j$ , then  $\hat{\kappa}$  does at least as well as  $\kappa$  for the loss  $\ell_v$ , up to sampling error (which is small by uniform convergence for intervals).
2. If  $\text{sign}(\kappa(p) - v)$  is not constant for bucket  $I_j$ , this means that the graph of  $\kappa(p)$  crosses the value  $v$  in this bucket, so the bucket is not a singleton.
  - By the definition of  $\mathcal{M}_r$ , this can happen for only  $2r$  buckets, which contain the endpoints of the  $r$  intervals that constitute the set  $\kappa(p) \geq v$ .
  - Since we use Uniform-mass bucketing, the total probability assigned to these buckets is small, which bounds how much worse  $\hat{\kappa}$  is than  $\kappa$ .



## Summary

Our work presents a comprehensive theory of the complexity of  $\text{CDL}_{\mathcal{K}}$  for binary classification. It delineates classes of post-processings (e.g. Lipschitz) that are intractable and classes that are efficient (e.g. generalized monotone functions). It proves a tight relationship to weight-restricted calibration. It introduces techniques from the literature on omniprediction for post-processing predictors, and proves rigorous new guarantees for some well-known algorithms used in practice.

We leave open the question of understanding efficient CDL for the multiclass setting where the number of labels  $k$  grows, as is the case in image classification. It is known that the problem in the weight-restricted setting becomes much harder for large  $k$ , indeed notions like smooth calibration and distance to calibration require sample complexity  $\exp(k)$  [GHR24]. It is a challenging question to formulate efficient and meaningful notions of CDL for this setting.

## 3 Preliminaries

In this section, we briefly review relevant concepts related to proper losses, calibration error, post-processing classes, and relevant fundamental concepts from learning theory. Proofs for results in this section appear in [Appendix B.1](#).

**Mathematical Miscellany** Given scalars  $t, a, b \in \mathbb{R}$  with  $a \leq b$ , we write  $[t]_a^b$  to denote the projection of  $t$  onto the closed interval  $[a, b]$ . When referring to an *interval*  $I \subseteq \mathbb{R}$  with no further specification, we mean that  $I$  may be open, closed, half-open, or a singleton. Phrased differently, an interval  $I$  may have the form  $[a, b]$ ,  $[a, b)$ ,  $(a, b]$ , or  $(a, b)$  for some  $a \leq b$ . This includes singleton sets of the form  $\{a\}$ . We shall sometimes write  $f(x) \lesssim g(x)$  to mean that  $f(x) = O(g(x))$ . Similarly,  $f(x) \gtrsim g(x)$  means  $f(x) = \Omega(g(x))$ . To avoid ambiguity, we will only use this notation when both functions  $f(x)$  and  $g(x)$  under consideration are nonnegative.

**Proper Losses** Since the class  $\mathcal{L}^*$  of proper losses will be our focus throughout this work, it will be instructive to review the following standard characterization of the functions that it contains.

**Lemma 3.1** (Proper Loss Characterization). *If  $\varphi : [0, 1] \rightarrow \mathbb{R}$  is concave and  $\varphi' : [0, 1] \rightarrow [-1, +1]$  is the derivative of  $\varphi$  (or an arbitrary superderivative if  $\varphi$  is nondifferentiable), then  $\mathcal{L}^*$  contains*

$$\ell(p, y) = \varphi(p) + \varphi'(p)(y - p).$$

*Conversely, every  $\ell \in \mathcal{L}^*$  has this form.*

When we first defined  $\mathcal{L}^*$ , we insisted that the *discrete partial derivative* with respect to the binary outcome, namely  $\partial\ell(p) = \ell(p, 1) - \ell(p, 0)$ , be bounded in absolute value (see [Definition 2.1](#)). The characterization in [Lemma 3.1](#) shows that this corresponds to a bound on  $\varphi'(p) = \partial\ell(p)$ . The characterization also leads to the following useful lemma, we call the *closer is better* lemma which states that shifting one's prediction toward the truth can only improve one's loss. To state it, recall that  $\ell(p, q) = \mathbb{E}_{y \sim \text{Ber}(q)}[\ell(p, y)]$ .

**Lemma 3.2** (Closer is better). *If  $\ell \in \mathcal{L}^*$  and  $a \leq b \leq c$ , then  $\ell(b, c) \leq \ell(a, c)$  and  $\ell(b, a) \leq \ell(c, a)$ .*

The preceding lemma can be used to show a bound on the range of any  $\ell \in \mathcal{L}^*$ :

**Lemma 3.3** (Bound on  $|\partial\ell|$  Implies Bound on  $|\ell|$ ). *If  $\ell \in \mathcal{L}^*$ , then there exists  $c \in \mathbb{R}$  such that*

$$c - 1 \leq \ell(p, y) \leq c + 1$$

*for all inputs  $(p, y) \in [0, 1] \times \{0, 1\}$ .*

Another consequence of the characterization in [Lemma 3.1](#) is the following description of the “boundary” of the convex set  $\mathcal{L}^*$ . Specifically, it is roughly the case that any loss in  $\mathcal{L}^*$  is a convex combination of a special type of proper loss functions, which we call *V-shaped* losses.

**Definition 3.4** (V-Shaped Loss). Given a scalar  $s \in [-1, +1]$ , consider the modified sign function

$$\text{sign}_s(t) = \begin{cases} +1 & \text{if } t > 0, \\ s & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

The *V-shaped* losses are the functions

$$\ell_{v,s}(p, y) = -\text{sign}_s(p - v)(y - v),$$

where  $v \in [0, 1]$  and  $s \in [-1, +1]$  are any values. Each loss  $\ell_{v,s}$  belongs to the class  $\mathcal{L}^*$ . We pay particular attention to the V-shaped losses  $\ell_v := \ell_{v,0}$  and  $\ell_v^+ := \ell_{v,(+1)}$ , and  $\ell_v^- := \ell_{v,(-1)}$ . We also write  $\text{sign}_+ = \text{sign}_{(+1)}$  and  $\text{sign}_- = \text{sign}_{(-1)}$  for brevity.

**Lemma 3.5** (Modification from [\[KLST23\]](#)). *Let  $\mathcal{L}_0$  be the set containing the V-shaped losses  $\ell_v^+$  and  $\ell_v^-$  for all  $v \in [0, 1]$ , as well as constant functions. Then for all  $\ell \in \mathcal{L}^*$  and  $\varepsilon > 0$ , there exists a convex combination  $\ell_0$  of finitely many functions in  $\mathcal{L}_0$  such that  $|\ell(p) - \ell_0(p)| \leq \varepsilon$  at all  $p \in [0, 1]$ .*

The preceding lemma can be used to reduce a search over all  $\ell \in \mathcal{L}^*$  to a search over a much smaller set. For example, in the original definition of CDL, it suffices to take the supremum over just the set of  $\ell_v^+$  and  $\ell_v^-$  losses, rather than all of  $\mathcal{L}^*$ . In fact, the following corollary further shows that it suffices to take the supremum over *just* the set of  $\ell_v^+$  losses *or* the set of  $\ell_v^-$  losses:

**Corollary 3.6.** *Given a distribution  $J$  over pairs  $(p, y)$ , post-processings  $\mathcal{K}$ , and  $\mathcal{L} \subseteq \mathcal{L}^*$ , let*

$$\text{CDL}_{\mathcal{L}, \mathcal{K}}(J) = \sup_{\ell, \kappa} \mathbb{E}[\ell(p, y) - \ell(\kappa(p), y)],$$

*where the supremum is taken over all  $\ell \in \mathcal{L}$  (not  $\mathcal{L}^*$ ) and  $\kappa \in \mathcal{K}$ . then*

$$\text{CDL}_{\mathcal{K}}(J) = \text{CDL}_{\mathcal{L}^+, \mathcal{K}}(J) = \text{CDL}_{\mathcal{L}^-, \mathcal{K}}(J),$$

*where  $\mathcal{L}^+ = \{\ell_v^+ \mid v \in [0, 1]\}$  and  $\mathcal{L}^- = \{\ell_v^- \mid v \in [0, 1]\}$ .*

In [Definition 2.9](#), we gave a broad definition of *weighted calibration error* measures, which is parameterized by a class of weight functions  $\mathcal{W}$ . The larger the class  $\mathcal{W}$ , the larger the corresponding weighted calibration error. When  $\mathcal{W}$  comprises all functions with range contained in  $[-1, +1]$ , we arrive at the standard notion of *expected calibration error* (ECE).

**Definition 3.7** (Expected Calibration Error). Given a distribution  $J$  over  $(p, y) \in [0, 1] \times \{0, 1\}$ ,

$$\text{ECE}(J) = \mathbb{E}[\mathbb{E}[y - p \mid p]].$$

As mentioned previously, [\[KLST23, HW24\]](#) showed that the calibration decision loss  $\text{CDL}_{\mathcal{K}^*}$  with respect to the class of *all* post-processing functions  $\mathcal{K}^* = \{[0, 1] \rightarrow [0, 1]\}$  is quadratically related to ECE. In particular, the upper bound also holds for  $\text{CDL}_{\mathcal{K}}$  for any subset  $\mathcal{K} \subseteq \mathcal{K}^*$ .

**Theorem 3.8** ([\[KLST23, HW24\]](#)). *Given a distribution  $J$  over  $(p, y) \in [0, 1] \times \{0, 1\}$ ,*

$$\text{ECE}(J)^2 \leq \text{CDL}_{\mathcal{K}^*}(J) \leq 2 \text{ECE}(J).$$



**Post-Processing** Here, we briefly discuss some terminology and facts related to post-processing classes that will be useful in subsequent sections. The first thing we need is a definition of what constitutes a *valid* post-processing class  $\mathcal{K}$ . For this, our only two requirements are that the  $\mathcal{K}$  contains the identity and is *translation invariant*.

**Definition 3.9** (Valid Post-Processing Class). We say that a class of functions  $\mathcal{K} \subseteq \{[0, 1] \rightarrow [0, 1]\}$  is a *valid post-processing class* if the following two conditions hold:

- **(Identity)**  $\mathcal{K}$  contains the identity function  $\kappa(p) = p$ ,
- **(Translation Invariance)** Fix any  $s, t \in \mathbb{R}$ . If  $\kappa$  belongs to  $\mathcal{K}$ , then so does the function

$$\kappa_{s,t}(p) = \left[ \kappa([p + s]_0^1) + t \right]_0^1.$$

The fact that  $\mathcal{K}$  contains the identity function means that  $\text{CDL}_{\mathcal{K}}$  will always be nonnegative. Later, we will show that the complexity of calibration decision loss relative to a valid post-processing classes  $\mathcal{K}$ , is closely related to the complexity of the class  $\text{thr}(\mathcal{K})$  of its upper thresholds.

**Definition 3.10** (Upper Thresholds). Given a valid post-processing class  $\mathcal{K}$ , its *upper thresholds* are

$$\text{thr}(\mathcal{K}) = \left\{ p \mapsto \text{sign}_+ \left( \kappa(p) - \frac{1}{2} \right) \mid \kappa \in \mathcal{K} \right\}.$$

Although the cutoff value of  $1/2$  in [Definition 3.10](#) may seem arbitrary, its choice is not especially important when working with valid post-processing classes, which are translation invariant.

**Definition 3.11** (VC Dimension). A class  $\mathcal{C} \subseteq \{[0, 1] \rightarrow \{\pm 1\}\}$  *shatters*  $S \subseteq [0, 1]$  if every function from  $S$  to  $\{\pm 1\}$  is the restriction of some function in  $\mathcal{C}$ . The *VC dimension* of  $\mathcal{C}$ , denoted  $\text{VCdim}(\mathcal{C})$ , is the size of the largest set it shatters. We say  $\text{VCdim}(\mathcal{C}) = \infty$  if the dimension is unbounded.

## 4 Sample Complexity of Testing and Auditing

In this section, we show that the sample complexity of testing/auditing is characterized (up to a quadratic gap) by the VC dimension of the class  $\text{thr}(\mathcal{K})$ . We use this to derive a lower bound for  $\text{CDL}_{\text{Lip}}$  where  $\text{Lip}$  denotes the family of 1-Lipschitz post-processings. We then consider the possibility of circumventing this lower bound by consider restricted loss families.

### 4.1 A Characterization via VC dimension

Our main sample complexity characterization is as follows.

**Theorem 4.1** (Sample Complexity Bounds). *Let  $\mathcal{K}$  be a valid post-processing class, and let  $\text{VCdim}(\text{thr}(\mathcal{K})) = d$ . Then,*

1. *For any  $\alpha, \varepsilon \in (0, 1)$ , there is an  $(\alpha, \alpha - \varepsilon)$ -tester for  $\text{CDL}_{\mathcal{K}}$  whose sample complexity is  $O(d \log(1/\varepsilon)/\varepsilon^2)$ .*
2. *Any  $(1/8, 0)$ -auditor for  $\text{CDL}_{\mathcal{K}}$  requires  $\Omega(\sqrt{d})$  samples.*

The upper bound relies on the following generalization result for CDL.

**Lemma 4.2** (Uniform Convergence for CDL). *Let  $\mathcal{K}$  be a valid post-processing class, and let  $\text{VCdim}(\text{thr}(\mathcal{K})) = d$ . For any  $\varepsilon, \delta \in (0, 1)$  and any set  $S$  of  $m = O(d \log(1/\varepsilon\delta)/\varepsilon^2)$  i.i.d. examples from some distribution  $J$  over  $[0, 1] \times \{0, 1\}$ , with probability at least  $1 - \delta$ , we have:*

$$\sup_{\substack{\ell \in \mathcal{L}^* \\ \kappa \in \mathcal{K}}} \left| \mathbb{E}_{(p,y) \sim J} [\ell(p, y) - \ell(\kappa(p), y)] - \mathbb{E}_{(p,y) \sim S} [\ell(p, y) - \ell(\kappa(p), y)] \right| \leq \varepsilon.$$

In particular, with probability at least  $1 - \delta$  we have  $|\text{CDL}_{\mathcal{K}}(J) - \text{CDL}_{\mathcal{K}}(S)| \leq \varepsilon$ .

Given this lemma, the upper bound is straightforward: we estimate  $\text{CDL}_{\mathcal{K}}(S)$  over a sufficiently large set of samples  $S$ , and decide to accept or reject by thresholding at  $\alpha - \varepsilon/2$ . We defer both the proof of [Lemma 4.2](#) and the derivation of the upper bound in [Theorem 4.1](#) to [Appendix B.2](#).

We now show the lower bound for auditing. Let  $p_1, p_2, \dots, p_d \in [0, 1]$  such that  $(\text{sign}_+(\kappa(p_i) - 1/2))_{i \in [d]}$  takes all the possible values in  $\{\pm 1\}^d$  for different choices of  $\kappa \in \mathcal{K}$ . By the translation invariance of  $\mathcal{K}$ , we can assume that there exist  $q_1, \dots, q_{d/2} \in [1/4, 3/4]$  such that  $(\text{sign}_+(\kappa(q_i) - 1/2))_{i \in [d/2]}$  takes all the possible values in  $\{\pm 1\}^{d/2}$  for different choices of  $\kappa \in \mathcal{K}$ .<sup>9</sup>

We consider the following distributions.

1. Let  $J_0$  be a distribution over  $[0, 1] \times \{0, 1\}$  whose marginal on  $[0, 1]$  is the uniform distribution over the set  $\{q_1, \dots, q_{d/2}\}$ , and  $\mathbb{E}_{(p,y) \sim J_0} [y|p = q_i] = q_i$ .
2. Let  $\mathcal{J}_1$  be a distribution over distributions that is defined as follows. To sample a distribution  $J_1$  from  $\mathcal{J}_1$ , we draw  $d/2$  independent random variables  $y_i \sim \text{Ber}(q_i)$  for  $i = 1, \dots, d/2$ . The marginal of  $J_1$  on  $[0, 1]$  is uniform over  $\{q_1, \dots, q_{d/2}\}$ , and  $\mathbb{E}_{(p,y) \sim J_1} [y|p = q_i] = y_i$ .

It is easy to show that  $J_0$  is perfectly calibrated, whereas every distribution  $J_1 \in \mathcal{J}_1$  has  $\text{ECE}(J_1) \geq 1/4$ . This is the basis of a lower bound for estimating the ECE in [\[GHR24\]](#). It is not true that  $\text{CDL}_{\mathcal{K}}(J_1)$  is large for every choice of  $J_1 \in \mathcal{J}_1$ . Nevertheless, we will show that  $\text{CDL}_{\mathcal{K}}(J_1)$  is likely to be large for a random  $J_1$ , which is sufficient for the lower bound to go through.

**Lemma 4.3.** *We have*

$$\Pr_{J_1 \sim \mathcal{J}_1} \left[ \text{CDL}_{\mathcal{K}}(J_1) \geq \frac{1}{8} \right] \geq \frac{5}{6}.$$

*Proof.* Consider the proper loss function

$$\ell(p, y) = \ell_{1/2}^+(p, y) = -(y - 1/2) \text{sign}_+(p - 1/2).$$

We will lower bound  $\text{CDL}_{\mathcal{K}}$  by

$$\begin{aligned} \text{CDL}_{\ell, \mathcal{K}}(J_1) &= \sup_{\kappa \in \mathcal{K}} \left( \frac{2}{d} \sum_{i=1}^{d/2} \left( y_i - \frac{1}{2} \right) \text{sign}_+ \left( \kappa(q_i) - \frac{1}{2} \right) - \text{sign}_+ \left( q_i - \frac{1}{2} \right) \right) \\ &= \underbrace{\sup_{\kappa \in \mathcal{K}} \left( \frac{2}{d} \sum_{i=1}^{d/2} \left( y_i - \frac{1}{2} \right) \text{sign}_+ \left( \kappa(q_i) - \frac{1}{2} \right) \right)}_{a_1} - \underbrace{\frac{2}{d} \sum_{i=1}^{d/2} \text{sign}_+ \left( q_i - \frac{1}{2} \right)}_{a_2} \end{aligned} \quad (4.1)$$

To lower bound  $a_1$ , we choose  $\kappa \in \mathcal{K}$  such that

$$\forall i, \text{sign}_+ \left( \kappa(q_i) - \frac{1}{2} \right) = \text{sign}_+ \left( y_i - \frac{1}{2} \right).$$

<sup>9</sup>If at least  $d/2$  points lie in  $[0, 1/2]$ , then we consider  $q = [p + 1/4]$  and  $\kappa'(q) = \kappa([q - 1/4]) = \kappa(p)$ . Else, at least  $d/2$  points lie in  $[1/2, 1]$ , so we consider  $q_i = [p_i - 1/4]$  and  $\kappa'(q) = \kappa([q + 1/4]) = \kappa(p_i)$ .

There exists such a  $\kappa$  for any realization of  $(y_i)_i$  due to the choice of  $(q_i)_i$ .<sup>10</sup> Since  $y_i \in \{0, 1\}$ , this ensures that

$$a_1 \geq \frac{2}{d} \sum_{i=1}^{d/2} \left( y_i - \frac{1}{2} \right) \text{sign}_+ \left( \kappa(q_i) - \frac{1}{2} \right) = \frac{2}{d} \sum_{i=1}^{d/2} \left| y_i - \frac{1}{2} \right| = \frac{1}{2}. \quad (4.2)$$

Next we show an upper bound on  $a_2$ . Over the random choice of  $y_i \sim \text{Ber}(q_i)$ , since  $\mathbb{E}[y_i|q_i] = q_i$ ,

$$\begin{aligned} \mathbb{E}[a_2] &= \frac{2}{d} \sum_{i=1}^{d/2} \left( q_i - \frac{1}{2} \right) \text{sign}_+ \left( q_i - \frac{1}{2} \right) \\ &= \frac{2}{d} \sum_{i=1}^{d/2} \left| q_i - \frac{1}{2} \right| \\ &\leq \frac{1}{4} \end{aligned}$$

where the last step uses the fact that  $q_i \in [1/4, 3/4]$ . By the Hoeffding bound, for  $d$  larger than some constant,  $\Pr[a_2 \geq 3/8] \leq 1/6$ . Therefore, with probability  $5/6$ , by Equations (4.1) and (4.2),

$$\text{CDL}_{\mathcal{K}}(J_1) \geq \frac{1}{2} - \frac{3}{8} = \frac{1}{8}.$$

□

We now complete the proof of the lower bound in [Theorem 4.1](#), using a birthday paradox argument, as in [\[GHR24\]](#).

*Proof of [Theorem 4.1](#), Lower bound.* Suppose that we have a  $(1/8, 0)$ -auditor  $\mathcal{A}$  for  $\text{CDL}_{\mathcal{K}}$  with sample complexity at most  $m$ . Since the distribution  $J_0$  is perfectly calibrated

$$\mathbb{P}_{S_0 \sim J_0^m} [\mathcal{A}(S_0) = \text{Accept}] \geq 2/3.$$

On the other hand, for  $J_1 \in \mathcal{J}_1$ , [Lemma 4.3](#) implies that  $\text{CDL}_{\mathcal{K}}(J_1) \geq 1/8$  with probability  $5/6$ . Conditioned on this event,  $\mathcal{A}$  being a  $(1/8, 0)$  auditor for  $\text{CDL}_{\mathcal{K}}$ ,  $\mathcal{A}(S_1)$  must reject with probability at least  $2/3$ . This means that

$$\mathbb{P}_{\substack{J_1 \sim \mathcal{J}_1 \\ S_1 \sim J_1^m}} [\mathcal{A}(S_1) = \text{Accept}] \leq \frac{1}{3} \cdot \frac{5}{6} + \frac{1}{6} < \frac{1}{2}. \quad (4.3)$$

But this means that the auditor  $\mathcal{A}$  satisfies the condition

$$\left| \mathbb{P}_{S_0 \sim J_0^m} [\mathcal{A}(S_0) = \text{Accept}] - \mathbb{P}_{\substack{J_1 \sim \mathcal{J}_1 \\ S_1 \sim J_1^m}} [\mathcal{A}(S_1) = \text{Accept}] \right| \geq 1/6, \quad (4.4)$$

where the probabilities are over the random variables  $S_0, J_1, S_1$  and any potential randomness of  $\mathcal{A}$ . If we condition on the events that all the elements of  $S_0$  and all the elements of  $S_1$  are distinct, then the corresponding conditional distributions are identical. Therefore, the total variation distance between the distribution of  $S_0$  and the distribution of  $S_1$  is bounded by the collision probability, which is smaller than  $1/6$  unless  $m = \Omega(\sqrt{d})$ . □

<sup>10</sup>This is in fact the  $\kappa$  that maximizes  $a_1$ .

**Lower bound for Lipschitz post-processings.** [Theorem 4.1](#) rules out efficient CDL estimation for the class of Lipschitz post-processings.

**Corollary 4.4.** *Let  $\text{Lip} \subset \mathcal{K}^*$  denote the class of 1-Lipschitz post-processings. There is no  $(1/8, 0)$ -auditor for  $\text{CDL}_{\text{Lip}}$  with finite sample complexity.*

This follows because  $\text{thr}(\text{Lip})$  has infinite VC dimension, this can be seen taking  $S$  to be a grid on the interval  $[0, 1]$  of multiples of  $2\gamma$ , and observing that the functions which take values in  $(1/2 \pm \gamma)$  on the grid points and interpolate linearly between them are Lipschitz. The thresholds of these functions shatter the set  $S$ . We now take  $\gamma \rightarrow 0$ .

**Upper bound for Generalized monotone post-processings.** [Theorem 4.1](#) implies a sample-efficient algorithm to estimate  $\text{CDL}_{\mathcal{M}_r}$  for the class  $\mathcal{M}_r$  of generalized monotone functions. This follows from [Proposition D.1](#) which bounds their VC dimension.

**Corollary 4.5.** *For all  $r \geq 1$ ,  $\alpha \in [0, 1]$  and  $\varepsilon < \alpha$ , there is an  $(\alpha, \alpha - \varepsilon)$ -tester for  $\text{CDL}_{\mathcal{M}_r}$  with sample complexity  $O(r \log(1/\varepsilon)/\varepsilon^2)$ .*

This bound by itself does not guarantee computational efficiency. But it can be made computationally efficient, as shown in [Corollary 6.2](#).

## 4.2 Other Families of Loss Functions

To conclude this section, we briefly discuss the role played by the class  $\mathcal{L}^*$  in our auditing lower bound. For example, one might ask whether our auditing lower bound continues to hold if we replace the class  $\mathcal{L}^*$  with natural subclasses that exclude the  $\ell_{1/2}^+$  loss, which played a key role in our proof. Here, we will investigate two such relaxations, corresponding to strongly proper loss functions, and loss functions that are Lipschitz in the predictions  $p$ . Note that V-shaped losses do not satisfy either of these conditions.

**Lower bounds for strongly proper losses.** We consider the class  $\mathcal{L}_{\mu\text{-sc}}^*$  of losses  $\ell \in \mathcal{L}^*$  whose associated concave function  $\varphi(p) = \mathbb{E}_{y \sim \text{Ber}(p)} \ell(p, y)$  satisfies  $\mu$ -strong concavity for some  $\mu > 0$ . To state the definition, recall that we write  $\ell(p, q) = \mathbb{E}_{y \sim \text{Ber}(q)} \ell(p, y)$ .

**Definition 4.6.** We say a function  $f : [0, 1] \rightarrow \mathbb{R}$  is  $\mu$ -strongly concave if for all  $x, y, \lambda \in [0, 1]$ ,

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) + \lambda(1 - \lambda) \cdot \frac{\mu}{2}(x - y)^2.$$

We let  $\mathcal{L}_{\mu\text{-sc}}^*$  denote the class of *strongly proper* losses  $\ell \in \mathcal{L}^*$ , meaning that  $\varphi(p) = \ell(p, p)$  is  $\mu$ -strongly concave in  $p$ . We let  $\text{CDL}_{\mathcal{L}_{\mu\text{-sc}}^*, \mathcal{K}}$  denote the similarly restricted version of  $\text{CDL}_{\mathcal{K}}$ .

While  $\varphi(p)$  is a concave function for any  $\ell \in \mathcal{L}^*$ , [Definition 4.6](#) goes beyond this by requiring *strong concavity*. For example,  $\varphi(p) = p(1 - p)$  is a  $\mu$ -strongly concave function with  $\mu = 2$ , corresponding to the squared loss  $\ell_{\text{sq}}(p, y) = (y - p)^2$ . We now study the testing and auditing of CDL with respect to this restricted class of proper losses, showing that the main results of this section continue to hold.

**Corollary 4.7** (Sample Complexity for  $\mathcal{L}_{\mu\text{-sc}}^*$ ). *Let  $\mathcal{K}$  be a valid post-processing class, and let  $\text{VCdim}(\text{thr}(\mathcal{K})) = d$ . Then, the following are true.*

1. For any  $\alpha, \varepsilon, \mu \in (0, 1)$ , there is an  $(\alpha, \alpha - \varepsilon)$ -tester for  $\text{CDL}_{\mathcal{L}_{\mu-\text{sc}}^*, \mathcal{K}}$  whose sample complexity is  $O(d \log(1/\varepsilon)/\varepsilon^2)$ .
2. For any  $\mu \in (0, 1/16)$ , any  $(1/8 - 2\mu, 0)$ -auditor for  $\text{CDL}_{\mathcal{L}_{\mu-\text{sc}}^*, \mathcal{K}}$  requires  $\Omega(\sqrt{d})$  samples.

We defer the proof to [Appendix B.2](#).

**Lipschitz losses and Smooth Calibration Error.** [Corollary 4.4](#) shows that  $\text{CDL}_{\mathcal{K}}$  is intractable when  $\mathcal{K}$  consists of Lipschitz post-processings. If we further restrict the proper losses to be Lipschitz, we will show that CDL is tightly characterized by the smooth calibration error [[KF08](#), [BGHN23a](#)], which is known to be both information-theoretically and computationally tractable.

We define our families of Lipschitz losses and post-processings.

**Definition 4.8.** Let  $\mathcal{L}_{\text{Lip}} \subset \mathcal{L}^*$  denote the family of all losses  $\ell$  such that  $\ell(p, y)$  is 1-Lipschitz in the first argument. Let  $\text{Lip}(2)$  denote the family of post-processing functions  $\kappa : [0, 1] \rightarrow [0, 1]$  that are 2-Lipschitz.

We consider 2-Lipschitz rather than 1-Lipschitz post-processings for technical reasons: in order to capture functions of the form  $[p + w(p)]_0^1$ , where  $w$  is 1-Lipschitz. This is convenient in order to provide a characterization of CDL in terms of  $\text{smCE}$ . But note that allowing more post-processings makes the positive result we will show more powerful, moreover the lower bound of [Corollary 4.4](#) also holds for the class of 2-Lipschitz post-processings.

Recall that the *smooth calibration error* is

$$\text{smCE}(J) = \sup_w \mathbb{E}[w(p)(y - p)],$$

where the supremum is taken over all  $w : [0, 1] \rightarrow [-1, 1]$  that are 1-Lipschitz.

**Theorem 4.9.** For any distribution  $J$  over  $[0, 1] \times \{0, 1\}$ , we have that:

$$\frac{1}{2}(\text{smCE}(J))^2 \leq \text{CDL}_{\mathcal{L}_{\text{Lip}}, \text{Lip}(2)}(J) \leq 6 \cdot \text{smCE}(J)$$

We present the proof in [Appendix B.2](#). The upper bound follows from the loss OI lemma of [[GHK<sup>+</sup>23](#)] ([Lemma 5.4](#)), while the lower bound follows from arguments in [[BGHN23b](#)]. Since  $\text{smCE}$  can be estimated efficiently from samples [[BGHN23a](#)], we have an efficient  $(\alpha, c\alpha^2)$ -auditor for  $\text{CDL}_{\mathcal{L}_{\text{Lip}}, \text{Lip}(2)}$ . We note that the restriction to Lipschitz losses is a significant one, which does exclude important losses. For instance, consider the  $\ell_1$  loss:  $\ell_1(p, y) = |p - y|$ . It is not proper, but it is Lipschitz in  $p$ . If we convert it to a proper loss by composing it with the best response, we get the  $\ell_{1/2}$  loss, which is not Lipschitz in  $p$ . This happens because the best-response  $\mathbf{1}[p \geq 1/2]$  is non-Lipschitz. This is not uncommon, even when the original loss is Lipschitz, since decision making in both theory and practice often involves sharp thresholds.

## 5 Relation to Weight-Restricted Calibration

In this section, we establish a tight relationship between  $\text{CDL}_{\mathcal{K}}$ , the calibration decision loss for a class  $\mathcal{K}$ , and a certain weight-restricted calibration error measure, a notion we defined in [Definition 2.9](#). The particular weight class we consider will involve thresholds of  $\mathcal{K}$ .

## 5.1 The Characterization

In order to state our general characterization, recall from [Definition 3.10](#) that  $\text{thr}(\mathcal{K})$  denotes the class of *upper thresholds* of a valid post-processing class  $\mathcal{K}$ . In this section, we will require a slight modification to the class  $\text{thr}(\mathcal{K})$ :

**Definition 5.1** (Modified Thresholds). Given a valid post-processing class  $\mathcal{K}$ , let

$$\text{thr}'(\mathcal{K}) = \text{thr}(\mathcal{K}) \cup \left\{ p \mapsto -\text{sign}_+(p - v) \mid v \in \mathbb{R} \right\}.$$

While the class  $\text{thr}(\mathcal{K})$  contains only the *upper* thresholds of functions in  $\mathcal{K}$ , the class  $\text{thr}'(\mathcal{K})$  also includes all *lower* thresholds of the identity function. It also contains upper thresholds of the identity function due to its inclusion of  $\text{thr}(\mathcal{K})$ , since  $\mathcal{K}$  is assumed to be translation invariant and contain the identity. The main result of this section is as follows:

**Theorem 5.2.** *If  $\mathcal{K}$  is a valid post-processing class, then*

$$\text{CE}_{\text{thr}'(\mathcal{K})}(J)^2 \lesssim \text{CDL}_{\mathcal{K}}(J) \lesssim \text{CE}_{\text{thr}'(\mathcal{K})}(J).$$

Before discussing the proof of [Theorem 5.2](#), we first present a couple of examples illustrating its use. Consider  $\mathcal{K} = \mathcal{M}_+$ , the class of monotonically nondecreasing post-processing functions. In this setting, the theorem implies that  $\text{CDL}_{\mathcal{M}_+}$  corresponds (up to quadratic equivalence) to the Interval-restricted calibration error  $\text{CE}_{\text{Int}}$  [[OKK25](#), [RSB<sup>+</sup>25](#)]*—*whose weight class consists of all upper and lower thresholds of the identity:  $\text{sign}(p - v)$  and  $-\text{sign}(p - v)$  for all  $v \in [0, 1]$ . The upper bound  $\text{CDL}_{\mathcal{M}_+}(J) \lesssim \text{CE}_{\text{Int}}(J)$  recovers a previous result from [[RSB<sup>+</sup>25](#)]. The lower bound  $\text{CDL}_{\mathcal{M}_+}(J) \gtrsim \text{CE}_{\text{Int}}(J)^2$ , however, is new. Moreover, [Theorem 5.2](#) is tight: there exist prediction-outcome distributions  $J_1$  and  $J_2$  such that  $\text{CDL}_{\mathcal{M}_+}(J_1) \lesssim \text{CE}_{\text{Int}}(J_1)^2$  and  $\text{CDL}_{\mathcal{M}_+}(J_2) \gtrsim \text{CE}_{\text{Int}}(J_2)$ . We defer discussion of these examples to [Appendix C](#).

Similarly, given  $r \in \mathbb{N}$ , we may set  $\mathcal{K} = \mathcal{M}_r$ , where  $\mathcal{M}_r$  is the class of  $r$ -wise generalized monotone post-processing functions defined in [Definition 2.2](#). In this case, [Theorem 5.2](#) implies that  $\text{CDL}_{\mathcal{M}_r}$  is quadratically related to a natural  $r$ -wise generalization of Interval-restricted calibration error, in which the weight class comprises all indicators for unions of at most  $r$  disjoint intervals.

The proof of [Theorem 5.2](#) will rely on the following few useful lemmas. The first lemma that we need is an immediate consequence of the bounded convergence theorem. It allows us to work with either strict or weak inequalities interchangeably, since  $\lim_{s \rightarrow t^+} \mathbf{1}[p \geq s] = \mathbf{1}[p > t]$ , etc.

**Lemma 5.3.** *Fix any distribution  $J$  over pairs  $(p, y) \in [0, 1] \times \{0, 1\}$ . Then, any pointwise convergent sequence of weight functions  $w_1, w_2, \dots : [0, 1] \rightarrow [-1, 1]$  satisfies*

$$\lim_{k \rightarrow \infty} \mathbb{E}[w_k(p)(y - p)] = \mathbb{E}\left[\lim_{k \rightarrow \infty} w_k(p)(y - p)\right].$$

The second lemma we need relates a gap in loss values to an expression that resembles weight-restricted calibration. We call it the *decision OI* lemma for its connection to *outcome indistinguishability (OI)* [[DKR<sup>+</sup>21](#)]*—*specifically, the notion of decision OI from [[GHK<sup>+</sup>23](#)].

**Lemma 5.4** (Decision OI). *For any fixed  $p, q \in [0, 1]$  and  $y \in \{0, 1\}$ ,*

$$\ell(p, y) - \ell(q, y) \leq (\partial \ell(p) - \partial \ell(q))(y - p).$$

*Proof.* Let  $\varphi$  be the concave function given by [Lemma 3.1](#) such that  $\ell(p, y) = \varphi(p) + \varphi'(p)(y - p)$ , where  $\varphi' = \partial\ell$  is some superderivative of  $\varphi$ . Then,

$$\ell(p, y) - \ell(q, y) = \underbrace{\varphi(p) - \varphi(q) - \varphi'(q)(p - q)}_{(*)} + (\varphi'(p) - \varphi'(q))(y - p),$$

and  $(*)$  is nonpositive by concavity.  $\square$

The third and final lemma we require is more subtle, providing a key relationship between interval miscalibration and the ability of an additive post-processing function to reduce loss. We call it the *small interval lemma*, since we will always apply it to intervals  $I$  with either short length  $b - a$ , or low mass  $\Pr[p \in I]$ .

**Lemma 5.5** (Small Interval). *If  $(p, y) \sim J$  and  $\mathcal{K}$  contains all functions  $\kappa(p) = [p + t]_0^1$  for  $t \in \mathbb{R}$ , then for all  $0 \leq a \leq b \leq 1$  and intervals  $I \in \{[a, b], [a, b), (a, b], (a, b)\}$ ,*

$$\left| \mathbb{E}[\mathbf{1}[p \in I](y - p)] \right| \leq (b - a) \Pr[p \in I] + \frac{1}{2} \text{CDL}_{\mathcal{K}}(J).$$

*Proof.* By the definition of calibration decision loss,

$$\text{CDL}_{\mathcal{K}}(J) \geq \sup_{v, t} \mathbb{E}[\ell_v^+(p, y) - \ell_v^+(\kappa_t(p), y)],$$

where the supremum is taken over all  $v \in [0, 1]$  and  $t \in [-1, 1]$ . This expectation equals

$$\mathbb{E}\left[\left(\text{sign}_+([p + t]_0^1 - v) - \text{sign}_+(p - v)\right)(y - v)\right] = \begin{cases} 2 \cdot \mathbb{E}[\mathbf{1}[v - t \leq p < v](y - v)] & \text{if } t > 0, \\ 2 \cdot \mathbb{E}[\mathbf{1}[v \leq p < v - t](v - y)] & \text{if } t < 0, \\ 0 & \text{if } t = 0. \end{cases}$$

Now consider any  $0 \leq a < b \leq 1$ . Taking  $v = b$  and  $t = b - a > 0$ , we see that

$$\mathbb{E}[\mathbf{1}[a \leq p < b](y - b)] \leq \frac{1}{2} \text{CDL}_{\mathcal{K}}(J). \quad (5.1)$$

Similarly, taking  $v = a$  and  $t = a - b < 0$ , we see that

$$\mathbb{E}[\mathbf{1}[a \leq p < b](a - y)] \leq \frac{1}{2} \text{CDL}_{\mathcal{K}}(J). \quad (5.2)$$

We visualize equations (5.1) and (5.2) in [Figure 2](#). Combining (5.1) and (5.2) yields

$$a \cdot \Pr[a \leq p < b] - \frac{1}{2} \text{CDL}_{\mathcal{K}}(J) \leq \mathbb{E}[y \cdot \mathbf{1}[a \leq p < b]] \leq b \cdot \Pr[a \leq p < b] + \frac{1}{2} \text{CDL}_{\mathcal{K}}(J).$$

Next, we subtract the quantity  $(a + b)/2 \cdot \Pr[a \leq p < b]$  from the left, right, and middle of the above chain of inequalities. Doing so yields

$$\left| \mathbb{E}\left[\mathbf{1}[a \leq p < b]\left(y - \frac{a + b}{2}\right)\right] \right| \leq \frac{b - a}{2} \Pr[a \leq p < b] + \frac{1}{2} \text{CDL}_{\mathcal{K}}(J).$$

Observe that, conditional on the event that  $a \leq p < b$ , the midpoint  $(a + b)/2$  of the interval  $[a, b]$  differs from the prediction  $p$  by at most  $(b - a)/2$ . Consequently, we have

$$\left| \mathbb{E}[\mathbf{1}[a \leq p < b](y - p)] \right| \leq (b - a) \Pr[a \leq p < b] + \frac{1}{2} \text{CDL}_{\mathcal{K}}(J).$$



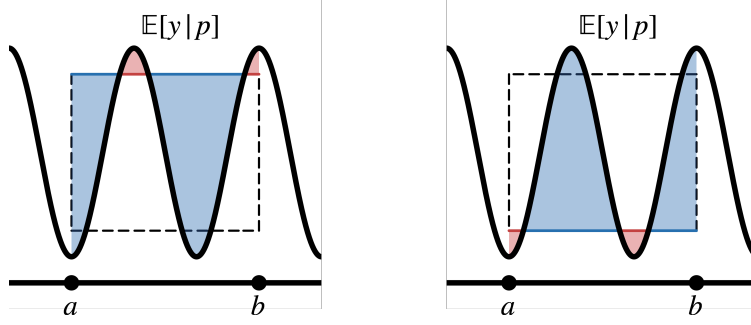


Figure 2: Visualization of inequalities (5.1) and (5.2). The dashed line outlines the box  $[a, b] \times [a, b]$  in the  $(p, y)$ -plane. The left and right images correspond to (5.1) and (5.2), respectively. The inequalities assert that in each image, the blue area is at least as large as the red, minus a  $\text{CDL}_{\mathcal{K}}(J)$  slack term, assuming the marginal distribution of  $p$  is uniform.

The left hand side is the absolute value of the calibration error restricted to the indicator function for the half-open interval  $I = [a, b)$ . Moreover, our argument up to this point works for any  $0 \leq a < b \leq 1$ , so that we have proved the claim for any such interval.

By carrying out the preceding argument with  $\ell_v^-$  in place of  $\ell_v^+$ , we can similarly bound the absolute calibration error restricted to half-open intervals of the form  $I = (a, b]$ , where  $0 \leq a < b \leq 1$ . By varying  $a$  and  $b$  and applying Lemma 5.3, we can also obtain the bound for any interval  $I$  of the form  $(a, b]$ ,  $(a, b)$ , or  $[a, b]$ , where  $0 \leq a \leq b \leq 1$ , except for  $I = [0, 1]$ , but this case is trivial because the right hand side is at least 1.  $\square$

With these three lemmas in hand, we are ready to prove Theorem 5.2.

*Proof of Theorem 5.2, Upper Bound.* As a reminder, our goal is to show that if  $\mathcal{K}$  is a valid post-processing class, then

$$\text{CDL}_{\mathcal{K}}(J) \lesssim \text{CE}_{\text{thr}'(\mathcal{K})}(J).$$

We first apply the boundary characterization of  $\mathcal{L}^*$  from Corollary 3.6, which yields the following upper bound:

$$\text{CDL}_{\mathcal{K}}(J) \leq \sup_{v, \kappa} \mathbb{E}[\ell_v^+(p, y) - \ell_v^+(\kappa(p), y)],$$

where the supremum is over  $v \in [0, 1]$  and  $\kappa \in \mathcal{K}$ . Next, we relate the gap in loss values between  $p$  and  $\kappa(p)$  to an expression resembling weight-restricted calibration using the decision OI lemma (Lemma 5.4):

$$\text{CDL}_{\mathcal{K}}(J) \leq \sup_{v, \kappa} \mathbb{E}[(\text{sign}_+(\kappa(p) - v) - \text{sign}_+(p - v))(y - p)],$$

By assumption, both  $\text{sign}_+(\kappa(p) - v)$  and  $-\text{sign}_+(p - v)$  belong to  $\text{thr}'(\mathcal{K})$ . We conclude that

$$\text{CDL}_{\mathcal{K}}(J) \leq 2 \cdot \text{CE}_{\text{thr}'(\mathcal{K})}(J). \quad \square$$

*Proof of Theorem 5.2, Lower Bound.* As a reminder, our goal is to show that if  $\mathcal{K}$  is a valid post-processing class, then

$$\text{CE}_{\text{thr}'(\mathcal{K})}(J)^2 \lesssim \text{CDL}_{\mathcal{K}}(J).$$

Suppose that  $\text{CDL}_{\mathcal{K}}(J) = \varepsilon$ . We will prove that  $\text{CE}_{\text{thr}'(\mathcal{K})}(J) \lesssim \sqrt{\varepsilon}$ . For this, we first recall that there are two types of weight functions in  $\text{thr}'(\mathcal{K})$ . The first kind of weight functions are functions of the form

$$w_1(p) = \text{sign}_+(\kappa(p) - v) = \mathbf{1}[\kappa(p) \geq v] \cdot 2 - 1,$$



for some  $\kappa \in \mathcal{K}$  and  $v \in [0, 1]$ . The second kind are functions of the form

$$w_2(p) = -\text{sign}_+(p - v) = \mathbf{1}[p < v] \cdot 2 - 1,$$

for some  $v \in [0, 1]$ . Therefore, it suffices to prove that the (signed) calibration errors with respect to all weight functions of the form  $p \mapsto \mathbf{1}[\kappa(p) \geq v]$  or  $p \mapsto \mathbf{1}[p < b]$  and  $p \mapsto -1$  are either negative, or positive but small (i.e.  $O(\sqrt{\varepsilon})$ ). We shall do so in the following two lemmas, [Lemmas 5.6](#) and [5.7](#). When combined, these two lemmas complete the proof of [Theorem 5.2](#).  $\square$

**Lemma 5.6.** *Let  $J$  be a distribution over pairs  $(p, y) \in [0, 1] \times \{0, 1\}$ , and let  $\mathcal{K}$  be a valid post-processing class. There exists an absolute constant  $C > 0$  such that for all  $\kappa \in \mathcal{K}$  and  $v \in [0, 1]$ ,*

$$\mathbb{E}[\mathbf{1}[\kappa(p) \geq v](y - p)] \leq C\sqrt{\text{CDL}_{\mathcal{K}}(J)}.$$

(Note that the expectation can be negative, with arbitrarily large magnitude.)

**Lemma 5.7.** *Let  $J$  be a distribution over pairs  $(p, y) \in [0, 1] \times \{0, 1\}$ , and let  $\mathcal{K}$  be a valid post-processing class. There exists an absolute constant  $C > 0$  such that for all intervals  $I \subseteq [0, 1]$ ,*

$$\left| \mathbb{E}[\mathbf{1}[p \in I](y - p)] \right| \leq C\sqrt{\text{CDL}_{\mathcal{K}}(J)}.$$

Of the two lemmas, [Lemma 5.7](#) has the slightly simpler proof, so we present it first.

*Proof of Lemma 5.7.* Let  $\varepsilon = \text{CDL}_{\mathcal{K}}(J)$  and let  $I \subseteq [0, 1]$  be an interval, which, as usual, may be open, closed, half-open, or a singleton. Let  $m \in \mathbb{N}$  be an integer to be specified later. We select a sequence of *cutoff* points  $c_1 < \dots < c_m$  such that  $c_1$  and  $c_m$  are the left and right endpoints of the interval  $I$ , and each open interval  $(c_{i-1}, c_i)$  contains at most a  $1/m$  fraction of the total probability mass of  $I$ . That is, we require

$$\Pr[c_{i-1} < p < c_i] \leq \frac{1}{m} \Pr[p \in I].$$

Because we consider open intervals  $(c_{i-1}, c_i)$ , such a selection is always possible. Note, however, that the distribution of  $p$  may have positive mass on some or all of the cutoff points  $c_1, \dots, c_m$ . Next, we split the associated weight-restricted calibration error at the cutoff points  $c_i$ , as follows:

$$\mathbb{E}[\mathbf{1}[p \in I](y - p)] = \sum_{i=1}^m \mathbb{E}[\mathbf{1}[p = c_i](y - p)] + \sum_{i=2}^m \mathbb{E}[\mathbf{1}[c_{i-1} < p < c_i](y - p)]. \quad (5.3)$$

By the small interval lemma ([Lemma 5.5](#)) each term the first sum in equation (5.3) is bounded in absolute value by  $O(\varepsilon)$ , and the  $i^{\text{th}}$  term in the second sum is bounded in absolute value by  $(c_i - c_{i-1}) \cdot 1/m + \varepsilon$ . Note that the difference  $c_i - c_{i-1}$  telescopes when we sum over  $i$ . Therefore, applying the triangle inequality to equation (5.3), we have that

$$\left| \mathbb{E}[\mathbf{1}[p \in I](y - p)] \right| \leq m\varepsilon + (c_m - c_1) \cdot \frac{1}{m} + m\varepsilon.$$

Recall that  $c_m - c_1$  is the length of  $I$ , which is at most 1. Setting  $m = \Theta(1/\sqrt{\varepsilon})$ , the above bound reduces to  $O(\sqrt{\varepsilon})$ , as claimed.  $\square$

Next, we present the proof of [Lemma 5.6](#), which has a similar structure to that of [Lemma 5.7](#), but with a few extra steps. This will conclude the proof of [Theorem 5.2](#).

*Proof of Lemma 5.6.* Once again, let  $\varepsilon = \text{CDL}_{\mathcal{K}}(J)$ . Instead of considering an interval, we now consider a set of the form  $S = \{p \in [0, 1] \mid \kappa(p) \geq v\}$  for some post-processing  $\kappa \in \mathcal{K}$  and value  $v \in [0, 1]$ . We again select cutoff points  $c_0 < \dots < c_m$  such that  $c_0 = 0$  and  $c_m = 1$  and each open interval  $(c_{i-1}, c_i)$  contains at most a  $1/m$  fraction of the mass of  $S$ . That is, we require  $\Pr[p \in S \cap (c_{i-1}, c_i)] \leq (1/m) \Pr[p \in S]$ . Since we consider open intervals  $(c_{i-1}, c_i)$ , such a selection is always possible. Note, however, that the distribution of  $p$  may have positive mass on some or all of the cutoff points  $c_0, \dots, c_m$ . Next, we split the associated weight-restricted calibration error at the cutoff points  $c_i$ , as follows:

$$\mathbb{E}[\mathbf{1}[p \in S](y - p)] = \sum_{i=0}^m \mathbb{E}[\mathbf{1}[p \in S \cap \{c_i\}](y - p)] + \sum_{i=1}^m \mathbb{E}[\mathbf{1}[p \in S \cap (c_{i-1}, c_i)](y - p)]. \quad (5.4)$$

Our goal will be to argue that each term of the two sums is either negative (with unbounded magnitude) or positive but small (i.e. with magnitude at most  $O(\sqrt{\varepsilon})$ ). First, by the small interval lemma (Lemma 5.5), the first sum in equation (5.4) is bounded in absolute value by  $O(m\varepsilon)$ . To bound the  $i^{\text{th}}$  term in the second sum, let  $a = c_{i-1}$  and  $b = c_i$ . We will consider the following three comprehensive cases, showing that the  $i^{\text{th}}$  term in the second sum of equation (5.4) is at most  $O((b - a)/m + \varepsilon)$ .

- **(Case 1:  $S \cap (a, b)$  is an interval)** In this case, Lemma 5.5 tells us that

$$\left| \mathbb{E}[\mathbf{1}[p \in S \cap (a, b)](y - p)] \right| \leq (b - a) \Pr[p \in S \cap (a, b)] + \varepsilon,$$

and the right hand side is at most  $(b - a)/m + \varepsilon$ .

- **(Case 2:  $S \cap (a, b)$  is not an interval and  $a \notin S$  and  $b \in S$ )** We first relate the calibration error on  $S \cap (a, b)$  to a quantity in which  $y - p$  has been replaced with  $y - b$ , as follows:

$$\mathbb{E}[\mathbf{1}[p \in S \cap (a, b)](y - p)] \leq \mathbb{E}[\mathbf{1}[p \in S \cap (a, b)](y - b)] + (b - a) \Pr[p \in S \cap (a, b)],$$

and the rightmost term is at most  $(b - a)/m$ . To bound the first term on the right, which involves a factor of  $y - b$ , we will relate the entire expectation to the  $\ell_b^+$  loss. To do so, recall that  $S = \{p \in [0, 1] \mid \kappa(p) \geq v\}$  for some  $\kappa \in \mathcal{K}$  and  $v \in [0, 1]$ . By translation invariance,  $\mathcal{K}$  also contains the function

$$\kappa'(p) = \left[ \kappa([p]_a^b) + b - v \right]_0^1.$$

We visualize the transformation from  $\kappa$  to  $\kappa'$  in Figure 3. Indeed, by shifting the graph of  $\kappa$  left by  $a$  (and back again) and right by  $1 - b$  (and back again), we can ensure that  $\kappa$  takes the constant value  $\kappa(a)$  on  $[0, a]$  and the constant value  $\kappa(b)$  on  $[b, 1]$ , which corresponds to the truncation  $[p]_a^b$  in the formula for  $\kappa'$ .

Next, we claim that  $\kappa'(p) \geq b$  if and only if  $p \in (S \cap (a, b)) \cup [b, 1]$ . To prove this, recall our assumption that  $a \notin S$ , which implies  $\kappa'(p) < b$  for all  $p \leq a$ . Similarly, since  $b \in S$ , we have  $\kappa'(p) \geq b$  for all  $p \geq b$ . For  $p \in (a, b)$ , we have  $\kappa'(p) \geq b$  if and only if  $p \in S$  as well. In terms of  $\kappa'$ , we have

$$\mathbb{E}[\mathbf{1}[p \in S \cap (a, b)](y - b)] = \mathbb{E}[(\mathbf{1}[\kappa'(p) \geq b] - \mathbf{1}[p \geq b])(y - b)].$$

This can be rewritten in terms of the  $\ell_v^+$  loss as

$$\mathbb{E}[\ell_b^+(p, y) - \ell_b^+(\kappa(p), y)].$$

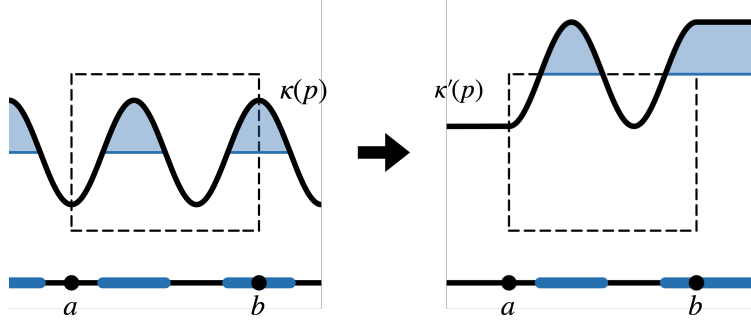


Figure 3: Transformation from  $\kappa$  to  $\kappa'$ . The dashed line outlines the box  $[a, b] \times [a, b]$  in the  $(p, \kappa(p))$ -plane. Starting from the set  $S = \{p \in [0, 1] : \kappa(p) \geq v\}$ , the transformation removes from  $S$  any points to the left of  $a$ , adds to  $S$  any points to the right of  $b$ , and performs a vertical shift so that the threshold coincides with  $b$ . The transformation requires  $\kappa(a) < v$  and  $\kappa(b) \geq v$ .

By the definition of  $\text{CDL}_{\mathcal{K}}$ , this quantity is either negative or at most  $O(\varepsilon)$ . We conclude that

$$\mathbb{E}[\mathbf{1}[p \in S \cap (a, b)](y - p)] \leq \frac{b - a}{m} + \varepsilon.$$

- **(Case 3:  $S \cap (a, b)$  is not an interval and either  $a \in S$  or  $b \notin S$ )** In this case, we will perform a simple trick to reduce Case 2, where  $a \notin S$  and  $b \in S$ . Basically, the idea is to shift  $a$  to the right until we hit a point not in  $S$ , and shift  $b$  to the left until we hit a point in  $S$ .

More formally, let  $a' = \inf(a, b) \setminus S$  be the infimum of points outside  $S$  in the interval, and let  $b' = \sup(a, b) \cap S$  be the supremum of points in  $S$  in the interval. Clearly,  $a' \geq a$  and  $b' \leq b$ . Also, we have the strict inequality  $a' < b'$  since  $S \cap (a, b)$  is not an interval.

Next, take any nonincreasing sequence of points  $a_j \notin S$  that approach  $a$ , as well as a nondecreasing sequence of points  $b_j \in S$  that approach  $b$ . Suppose also that we have the strict inequality  $a_j < b_j$ . To bound the calibration error restricted to  $S \cap (a', b')$ , we apply the result of Case 2 to each pair  $(a_j, b_j)$  and take the limit as  $j \rightarrow \infty$  (Lemma 5.3). This yields

$$\mathbb{E}[\mathbf{1}[p \in S \cap (a', b')](y - p)] \leq \frac{b' - a'}{m} + \varepsilon.$$

The calibration errors restricted to  $S \cap (a, a')$  and  $S \cap [b', b]$  are similarly bounded by the small interval lemma (Lemma 5.5), since the former is a contiguous interval and the latter is either a singleton or empty set. This yields

$$\mathbb{E}[\mathbf{1}[p \in S \cap (a, a')](y - p)] \leq \frac{a' - a}{m} + \varepsilon$$

and

$$\mathbb{E}[\mathbf{1}[p \in S \cap (b', b)](y - p)] \leq \frac{b - b'}{m} + \varepsilon.$$

Adding these three inequalities, we see that, in total, over  $S \cap (a, a')$ ,  $S \cap (a', b')$ , and  $S \cap [b', b]$ , we have

$$\mathbb{E}[\mathbf{1}[p \in S \cap (a, b)](y - p)] \leq \frac{b - a}{m} + 3\varepsilon.$$

At this point, we have shown that the  $i^{\text{th}}$  term of the second summation of equation (5.4) is either negative (with arbitrarily large magnitude) or positive but bounded in magnitude by  $O((c_i - c_{i-1})/m + \varepsilon)$ . Setting  $m = \Theta(1/\sqrt{\varepsilon})$ , the entire bound telescopes to  $O(\sqrt{\varepsilon})$ , proving Lemma 5.6.  $\square$

## 5.2 Translation Invariance Alone Is Insufficient

We have established that  $\text{CDL}_{\mathcal{K}}$  is polynomially characterized by the  $\text{thr}'(\mathcal{K})$ -restricted calibration error for any valid post-processing class  $\mathcal{K}$ . Here, we provide an example that justifies these assumptions, by showing that the characterization does not hold for the class of non-increasing post-processings. Recall that the class of non-increasing post-processings is translation invariant, but does not contain the identity.

**Theorem 5.8.** *Let  $\mathcal{K} = \mathcal{M}_-$  and  $\mathcal{W} = \text{thr}'(\mathcal{M}_-)$ . There is a distribution  $J$  such that:*

$$\text{CE}_{\mathcal{W}}(J) \geq \frac{1}{4}, \text{ yet } \text{CDL}_{\mathcal{K}}(J) \leq 0$$

*Proof.* Let  $J$  be the distribution over  $(p, y)$  such that the marginal distribution of  $p$  is 0 with probability 1/2 and 1/2 with probability 1/2. Moreover, we have

$$p^*(p) := \mathbb{E}_{(p,y) \sim J}[y|p] = \begin{cases} 0, & \text{if } p = 0 \\ 1, & \text{if } p = 1/2 \end{cases}$$

The set  $\mathcal{W}$  contains the constant function  $w^* : p \mapsto 1$ , corresponding to the nonincreasing constant postprocessing  $\kappa : p \mapsto 1$ . Therefore, we have the following:

$$\begin{aligned} \text{CE}_{\mathcal{W}}(J) &\geq \mathbb{E}_{(p,y) \sim J}[w^*(p)(y - p)] \\ &= \mathbb{E}_{(p,y) \sim J}[y - p] \\ &= \frac{1}{2} \mathbb{P}_{(p,y) \sim J}[p = 1/2] = \frac{1}{4}. \end{aligned}$$

Due to [Corollary 3.6](#), in order to show that  $\text{CDL}_{\mathcal{K}}(J) \leq 0$ , it suffices to show that  $Q_{v,\kappa}^* \leq 0$  for all  $v \in [0, 1]$ ,  $\kappa \in \mathcal{K}$  and  $\star \in \{+, -\}$ , where  $Q_{v,\kappa}^*$  is defined as follows:

$$\begin{aligned} Q_{v,\kappa}^* &:= \mathbb{E}_{(p,y) \sim J}[(y - v)(\text{sign}_{\star}(\kappa(p) - v) - \text{sign}_{\star}(p - v))] \\ &= \frac{1}{2}(-v)(\text{sign}_{\star}(\kappa(0) - v) - \text{sign}_{\star}(-v)) + \frac{1}{2}(1 - v)(\text{sign}_{\star}(\kappa(1/2) - v) - \text{sign}_{\star}(1/2 - v)) \\ &= \frac{1}{2}(-v)(\text{sign}_{\star}(\kappa(0) - v) + 1) + \frac{1}{2}(1 - v)(\text{sign}_{\star}(\kappa(1/2) - v) - \text{sign}_{\star}(1/2 - v)) \end{aligned}$$

The first term in the above expression is non-positive. Therefore,  $Q_{v,\kappa}^* \leq 0$  unless the second term is positive. If the second term is positive, then  $\text{sign}_{\star}(\kappa(1/2) - v) = 1$ , which implies, due to monotonicity of  $\kappa$ , that  $\text{sign}_{\star}(\kappa(0) - v) = 1$ . Assuming that  $\text{sign}_{\star}(\kappa(1/2) - v) = \text{sign}_{\star}(\kappa(0) - v) = 1$ , we have the following cases.

**Case I:**  $v \geq 1/2$ . We have  $Q_{v,\kappa}^* \leq -v + (1 - v) = 1 - 2v \leq 0$ .

**Case II:**  $v < 1/2$ . We have  $Q_{v,\kappa}^* = -v \leq 0$ . □

*Remark 5.9.* The characterization of [Theorem 5.2](#) fails even if we consider a class  $\mathcal{K}$  that includes both  $\mathcal{M}_-$  and the identity function, since the identity function does not alter the loss value. Hence, neither translation invariance nor inclusion of the identity function alone suffices for the characterization of [Theorem 5.2](#).

## 6 Computationally Efficient Testing and Auditing

In this section, we show that the problem of testing  $\text{CDL}_{\mathcal{K}}$  for some post-processing class  $\mathcal{K}$  can be reduced to proper agnostic learning of the threshold class  $\text{thr}(\mathcal{K})$ , whereas auditing reduces to improper agnostic learning. Note that the characterization of calibration decision loss in terms of the calibration error with respect to the threshold class we provided in [Section 5](#) already establishes a connection between agnostic learning and CDL testing: it says that an agnostic learner for  $\text{thr}(\mathcal{K})$  gives an  $(\alpha, c\alpha^2 - \varepsilon)$  tester for some constant  $c$ . However, our testing guarantee here is stronger, since we get  $\beta = \alpha - \varepsilon$ .

**Theorem 6.1** (Testing and Auditing from Agnostic Learning). *Let  $\mathcal{K}$  be a valid post-processing class. Let  $\text{AL}$  be an agnostic  $\varepsilon$ -learner for  $\text{thr}(\mathcal{K})$  with sample complexity  $m$ . For any  $\alpha \in (0, 1)$ , there is an  $(\alpha, \alpha - 3\varepsilon)$ -auditor for  $\text{CDL}_{\mathcal{K}}$  that makes at most  $O((1/\varepsilon) \log(1/\varepsilon\delta))$  non-adaptive calls to  $\text{AL}$ , uses  $O((1/\varepsilon^2) \log(1/\varepsilon\delta))$  additional samples, and performs  $\tilde{O}(\log(1/\delta)/\varepsilon^3)$  additional operations. Moreover, if  $\text{AL}$  is proper, then there is an  $(\alpha, \alpha - 3\varepsilon)$ -tester for  $\text{CDL}_{\mathcal{K}}$  with the same specifications.*

We obtain the following corollary for the class of generalized monotone post-processings, based on a folklore result on agnostic learning unions of intervals in one dimension (see [Appendix D](#)).

**Corollary 6.2.** *For any  $r \geq 1$ ,  $\alpha, \varepsilon \in (0, 1)$ , there is an  $(\alpha, \alpha - \varepsilon)$ -tester for  $\text{CDL}_{\mathcal{M}_r}$  with sample complexity  $\tilde{O}(r/\varepsilon^2)$  and runtime  $\tilde{O}(r^2/\varepsilon^3)$ .*

The proof of [Theorem 6.1](#) is based on the characterization of proper losses in terms of the V-shaped losses ([Lemma 3.5](#)) and an appropriate discretization argument for the parameter  $v \in [0, 1]$  associated with the family of V-shaped losses. For each (discrete) choice of  $v$ , we identify a relevant agnostic learning problem and solve it using the agnostic learning oracle. First, we state the version of the result for a fixed choice of  $v$  separately, as a lemma.

**Lemma 6.3.** *Let  $\mathcal{K}$  be a valid post-processing class. Given  $v \in [0, 1]$ , denote the calibration fixed decision loss with respect to the loss  $\ell_v^+$  by*

$$\text{CDL}_{v, \mathcal{K}}(J) = \sup_{\kappa \in \mathcal{K}} \mathbb{E}[\ell_v^+(p, y) - \ell_v^+(\kappa(p), y)].$$

*Let  $\text{AL}$  be an agnostic  $(\varepsilon, \delta)$ -learner for  $\text{thr}(\mathcal{K})$  with sample complexity  $m$ . Then, for any  $\alpha, \gamma \in (0, 1)$ , there exists an algorithm that calls  $\text{AL}$  once, uses  $O((1/\gamma^2) \log(1/\delta))$  additional samples and operations, and with probability at least  $1 - \delta$  outputs an estimate  $\widehat{\text{CDL}}_{v, \mathcal{K}}(J)$  satisfying*

$$\widehat{\text{CDL}}_{v, \mathcal{K}}(J) \in [\text{CDL}_{v, \mathcal{K}}(J) - 2\varepsilon - 2\gamma, 2 \cdot \text{ECE}(J) + 2\gamma].$$

*Moreover, if  $\text{AL}$  is proper, then we have the stronger guarantee*

$$\widehat{\text{CDL}}_{v, \mathcal{K}}(J) \in [\text{CDL}_{v, \mathcal{K}}(J) - 2\varepsilon - 2\gamma, \text{CDL}_{v, \mathcal{K}}(J) + 2\gamma].$$

*Proof.* Our strategy will be to manipulate the expression for  $\text{CDL}_{v, \mathcal{K}}$  into a form amenable to agnostic learning. First, we substitute the definition of the loss function  $\ell_v^+(p, y) = -\text{sign}_+(p - v)(y - v)$ , and use the translation invariance of the class  $\mathcal{K}$ . This allows us to rewrite the formula for  $\text{CDL}_{v, \mathcal{K}}$  as

$$\text{CDL}_{v, \mathcal{K}}(J) = \mathbb{E}[-\text{sign}_+(p - v)(y - v)] + \sup_{\kappa \in \mathcal{K}} \mathbb{E}\left[\text{sign}_+\left(\kappa(p) - \frac{1}{2}\right)(y - v)\right].$$

By Hoeffding's inequality, the first expectation in the above equation can be directly estimated up to error  $\gamma$  with probability at least  $1 - \delta$  from a sample of  $(p, y)$  pairs of size  $O(\log(1/\delta)/\gamma^2)$ .

We can similarly get a handle on the second term, which includes a supremum over post-processing functions  $\kappa$ , using a single call to the agnostic learner with concept class  $\mathcal{C} = \text{thr}(\mathcal{K})$  and labels  $z = y - v \in [-1, +1]$ . (For discrete labels  $z \in \{\pm 1\}$ , simply perform randomized rounding that preserves  $z$  in expectation.) If the agnostic learner returns a hypothesis  $h : [0, 1] \rightarrow \{\pm 1\}$ , then we have the following guarantee with probability at least  $1 - \delta$ :

$$\mathbb{E}[h(p)(y - v)] \geq \sup_{\kappa \in \mathcal{K}} \mathbb{E}\left[\text{sign}_+\left(\kappa(p) - \frac{1}{2}\right)(y - v)\right] - 2\varepsilon.$$

Note also that  $h(p) = \text{sign}_+(\kappa(p) - 1/2)$  for some function  $\kappa : [0, 1] \rightarrow [0, 1]$ , which we may assume belongs to the class  $\mathcal{K}$  if the agnostic learner is *proper*. Consequently, we have

$$\mathbb{E}[h(p)(y - v)] \leq \sup_{\kappa} \mathbb{E}\left[\text{sign}_+\left(\kappa(p) - \frac{1}{2}\right)(y - v)\right],$$

where the supremum is taken over all  $\kappa : [0, 1] \rightarrow [0, 1]$  in case of an improper learner, or all  $\kappa \in \mathcal{K}$  in the case of a proper learner. Of course, given an additional  $O(\log(1/\delta)/\gamma^2)$  samples, we can estimate the quantity  $\mathbb{E}[h(p)(y - v)]$  up to error  $\gamma$  with probability at least  $1 - \delta$ .

At this point, our algorithm makes a single call to **AL**, uses  $O(\log(1/\delta)/\gamma^2)$  additional samples, and with probability at least  $1 - \delta$  outputs a scalar estimate  $\widehat{\text{CDL}}_{v, \mathcal{K}}(J)$  satisfying

$$\text{CDL}_{v, \mathcal{K}}(J) - 2\varepsilon - 2\gamma \leq \widehat{\text{CDL}}_{v, \mathcal{K}}(J) \leq \text{CDL}_{v, \mathcal{K}^*}(J) + 2\gamma,$$

where  $\mathcal{K}^*$  is the set of all post-processings  $\kappa : [0, 1] \rightarrow [0, 1]$ . By [Theorem 3.8](#),  $\text{CDL}_{v, \mathcal{K}^*}(J) \leq 2 \cdot \text{ECE}(J)$ . In the special case that **AL** is a proper agnostic learner, we have the following stronger condition, with  $\mathcal{K}$  replacing  $\mathcal{K}^*$  in the upper bound:

$$\text{CDL}_{v, \mathcal{K}}(J) - 2\varepsilon - 2\gamma \leq \widehat{\text{CDL}}_{v, \mathcal{K}}(J) \leq \text{CDL}_{v, \mathcal{K}}(J) + 2\gamma. \quad \square$$

Having shown that agnostic learning can be used to understand the calibration fixed decision loss with respect to a particular  $\ell_v^+$  loss, we now prove [Theorem 6.1](#), which shows that proper and improper agnostic learning similarly imply testing and auditing for  $\text{CDL}_{\mathcal{K}}$ .

*Proof of Theorem 6.1.* Suppose momentarily that [Lemma 6.3](#) were to hold *simultaneously* for all values  $v \in [0, 1]$ , rather than one fixed value. Then, the supremum of the estimates  $\widehat{\text{CDL}}_{v, \mathcal{K}}$  would lie between  $\text{CDL}_{\mathcal{K}}(J) - 2\varepsilon - 2\gamma$  and  $2 \cdot \text{ECE}_{\mathcal{K}}(J) + 2\gamma$ . In the special case that **AL** is proper, the supremum would also lie below  $\text{CDL}_{\mathcal{K}}(J) + 2\gamma$ . For  $\gamma < \varepsilon/4$ , this would immediately imply  $(\alpha, \alpha - 3\varepsilon)$ -auditing, or, in the case of a proper agnostic learner,  $(\alpha, \alpha - 3\varepsilon)$ -testing.

With this in mind, our strategy will be to efficiently obtain these estimates  $x_v$  for a sufficiently well-spaced *net* of points  $v_1, \dots, v_t \in [0, 1]$ . By “well-spaced,” we mean that replacing any  $v \in [0, 1]$  with its nearest neighbor in the net should change the value of  $\text{CDL}_{v, \mathcal{K}}(J)$  by at most  $O(\gamma)$ . For this, recall the formula for the calibration fixed decision loss with respect to  $\ell_v^+$ :

$$\text{CDL}_{v, \mathcal{K}}(J) = \mathbb{E}[-\text{sign}_+(p - v)(y - v)] + \sup_{\kappa \in \mathcal{K}} \mathbb{E}\left[\text{sign}_+\left(\kappa(p) - \frac{1}{2}\right)(y - v)\right].$$

By inspection of this formula, it is clear that it suffices for every  $v \in [0, 1]$  to have a nearest neighbor  $v_i$  in the net that is both *close in length* and *close in probability*:

- **(Close in Length)**  $|v - v_i| \leq \gamma$ ,
- **(Close in Probability)**  $\Pr[\text{sign}_+(p - v) \neq \text{sign}_+(p - v_i)] \leq \gamma$ .

Such a net can be constructed by taking the union of equally spaced points  $0, \gamma, 2\gamma, \dots, 1$  with several independent samples  $p_1, \dots, p_s$  drawn from the distribution  $J$ . The equally spaced points clearly guarantee closeness in length. Closeness in probability follows from the fact that we can partition the interval  $[0, 1]$  into  $O(1/\gamma)$  contiguous subintervals, each of which either has probability mass  $\Theta(\gamma)$  or is a singleton of mass  $\Omega(\gamma)$ . Then, as long as  $s = O(\log(1/\gamma\delta)/\gamma)$ , with probability at least  $1 - \delta$ , every subinterval contains some sampled point  $p_i$ .

To conclude, we observe that for  $\gamma = \Omega(\varepsilon)$ , calling our fixed- $v$  algorithm from [Lemma 6.3](#) as a subroutine for each point in the net leads to a total of  $O((1/\varepsilon) \log(1/\varepsilon\delta))$  calls. By a union bound over the  $\delta$  failure probability of each call to the subroutine, we need only  $O((1/\varepsilon^2) \log(1/\varepsilon\delta))$  additional samples, in total. Similarly, the total number of additional operations is also  $\tilde{O}(\log(1/\delta)/\varepsilon^3)$ .  $\square$

## 7 Omniprediction

Recall the definition of omniprediction ([Definition 2.11](#)). In this section, we prove the following omniprediction guarantees:

1. We show that if  $\text{thr}(\mathcal{K})$  is agnostically learnable, then one can efficiently learn an  $(\varepsilon, \mathcal{K})$  omnipredictor. This result is proved by adapting the loss OI framework of [\[GHK<sup>+</sup>23\]](#) to the calibration setting.
2. We provide a strong omniprediction guarantee for the classical Pool Adjacent Violators algorithm [\[ABE<sup>+</sup>55\]](#): it gives a proper omnipredictor for the class of all monotone post-processings.
3. We also provide an analysis of bucketed recalibration through uniform-mass binning, which has been studied in [\[ZE01, GR21, SSH23\]](#). We show that it gives an omnipredictor for generalized monotone functions.

### 7.1 Omniprediction From Agnostic Learning

We show a general result that reduces omniprediction for  $\mathcal{K}$  to agnostic learning for  $\text{thr}(\mathcal{K})$ . Thus under the same computational assumptions that we needed for efficient auditing of  $\text{CDL}_{\mathcal{K}}$ , we can get the strong post-processing guarantee of omniprediction.

The following theorem reduces omniprediction with respect to all proper losses and some post-processing class  $\mathcal{K}$  to agnostic learning of the class  $\text{thr}(\mathcal{K})$ .

**Theorem 7.1** (Omniprediction from Agnostic Learning). *Let  $\mathcal{K}$  be a valid post-processing class. Let  $\text{AL}$  be an  $(\varepsilon/3)$ -agnostic learner for  $\text{thr}(\mathcal{K})$ . There is an algorithm that learns an  $(\varepsilon, \mathcal{K})$ -omnipredictor with probability  $1 - \delta$  that calls  $\text{AL}$   $O(\log(1/\delta)/\varepsilon^2)$  times and performs  $\text{poly}(1/\varepsilon) \log(1/\delta)$  additional operations.*

[\[GHK<sup>+</sup>23\]](#) show that omniprediction follows from a condition called calibrated multiaccuracy. Below we define a version of calibrated multiaccuracy that is tailored to our application.

**Definition 7.2** (Calibrated Multiaccuracy). Let  $\hat{\kappa} : [0, 1] \rightarrow [0, 1]$  and  $\mathcal{C} \subseteq \{[0, 1] \rightarrow \{\pm 1\}\}$ . We say that  $\hat{\kappa}$  is  $\varepsilon$ -calibrated-multiaccurate w.r.t.  $\mathcal{C}$  under the distribution  $J$ , or  $\hat{\kappa} \in \text{calMA}_{\mathcal{C}}(J, \varepsilon)$ , if the following properties hold.

1. (Calibration).  $\text{ECE}(\hat{J}) \leq \varepsilon$ , where  $\hat{J}$  is the distribution of pairs  $(\hat{\kappa}(p), y)$ , where  $(p, y) \sim J$ .
2. (Multiaccuracy).  $|\mathbb{E}_{(p, y) \sim J}[c(p)(y - \hat{\kappa}(p))]| \leq \varepsilon$  for all  $c \in \mathcal{C}$ .



We now prove the following lemma, which states that calibrated multiaccuracy implies omniprediction.

**Lemma 7.3.** *Let  $\hat{\kappa} \in \text{calMA}_{\mathcal{C}}(J, \varepsilon)$ , where  $\mathcal{C} = \text{thr}(\mathcal{K})$ . Then  $\hat{\kappa}$  is a  $(2\varepsilon, \mathcal{K})$ -omnipredictor.*

*Proof.* We will show that for any  $\kappa \in \mathcal{K}$  and any  $\ell \in \mathcal{L}^*$  the following holds.

$$\mathbb{E}_{(p,y) \sim J}[\ell(\hat{\kappa}(p), y)] \leq \mathbb{E}_{(p,y) \sim J}[\ell(\kappa(p), y)] + 2\varepsilon \quad (7.1)$$

We define the distribution  $\tilde{J}$  to be the distribution  $(p, \tilde{y})$  where  $p$  has the same as the marginal distribution as under  $J$  and  $\tilde{y} \sim \text{Ber}(\hat{\kappa}(p))$  so that  $\mathbb{E}[\tilde{y}|p] = \hat{\kappa}(p)$ .

For the following, for any  $\ell \in \mathcal{L}^*$ , we let  $\partial\ell : [0, 1] \rightarrow [-1, 1]$  denote the function  $\partial\ell(p) = \ell(p, 1) - \ell(p, 0)$ . Note that for any  $p \in [0, 1]$  and  $y \in \{0, 1\}$ , we have  $\ell(p, y) = y\partial\ell(p) + \ell(p, 0)$ .

We will first bound the quantity  $\Delta_1 := \mathbb{E}_J[\ell(\hat{\kappa}(p), y)] - \mathbb{E}_{\tilde{J}}[\ell(\hat{\kappa}(p), \tilde{y})]$  for any  $\ell \in \mathcal{L}^*$  as follows.

$$\begin{aligned} \Delta_1 &= \mathbb{E}_J[y\partial\ell(\hat{\kappa}(p)) + \ell(\hat{\kappa}(p), 0)] - \mathbb{E}_{\tilde{J}}[\tilde{y}\partial\ell(\hat{\kappa}(p)) + \ell(\hat{\kappa}(p), 0)] \\ &= \mathbb{E}_J[y\partial\ell(\hat{\kappa}(p))] - \mathbb{E}_{\tilde{J}}[\tilde{y}\partial\ell(\hat{\kappa}(p))] \quad (J, \tilde{J} \text{ have the same } p\text{-marginal}) \\ &= \mathbb{E}_J[(y - \hat{\kappa}(p))\partial\ell(\hat{\kappa}(p))] \quad (\mathbb{E}[\tilde{y}|p] = \hat{\kappa}(p)) \\ &= \mathbb{E}_{(q,y) \sim \hat{J}}[(y - q)\partial\ell(q)] \quad ((\hat{\kappa}(p), y) \sim \hat{J}) \\ &\leq \sup_{w: [0,1] \rightarrow [-1,1]} \left| \mathbb{E}_{\hat{J}}[(y - q)w(q)] \right| = \text{ECE}(\hat{J}) \leq \varepsilon \quad (\text{ECE}(\hat{J}) \leq \varepsilon) \end{aligned}$$

On the other hand, since  $\ell$  is a proper loss and  $\mathbb{E}[\tilde{y}|p] = \hat{\kappa}(p)$ , [Definition 2.1](#) implies that for any  $p, p' \in [0, 1]$ ,  $\mathbb{E}_{\tilde{J}}[\ell(\hat{\kappa}(p), \tilde{y})|p] \leq \mathbb{E}_{\tilde{J}}[\ell(p', \tilde{y})|p]$ . It follows that for any  $\kappa \in \mathcal{K}$ ,  $\mathbb{E}_{\tilde{J}}[\ell(\hat{\kappa}(p), \tilde{y})] \leq \mathbb{E}_{\tilde{J}}[\ell(\kappa(p), \tilde{y})]$ . Combining this with the bound on  $\Delta_1$  yields

$$\mathbb{E}_{(p,y) \sim J}[\ell(\hat{\kappa}(p), y)] \leq \mathbb{E}_{(p,\tilde{y}) \sim \tilde{J}}[\ell(\kappa(p), \tilde{y})] + \varepsilon \quad (7.2)$$

We will now show that the quantity  $\Delta_2 := \mathbb{E}_{(p,\tilde{y}) \sim \tilde{J}}[\ell(\kappa(p), \tilde{y})] - \mathbb{E}_{(p,y) \sim J}[\ell(\kappa(p), y)]$  satisfies  $\Delta_2 \leq \varepsilon$  for any  $\ell \in \mathcal{L}^*$  and  $\kappa \in \mathcal{K}$ . Combining the bound on  $\Delta_2$  with Eq. (7.2) implies Eq. (7.1).

By [Lemma 3.5](#), we can write

$$\partial\ell(p) = - \int_{[0,1]} \text{sign}_+(p - v) d\mu_{\ell}^+(v) - \int_{[0,1]} \text{sign}_-(p - v) d\mu_{\ell}^-(v),$$

where  $\mu_{\ell}^{\pm}$  are measures over  $[0, 1]$  such that  $\mu_{\ell}^+([0, 1]) + \mu_{\ell}^-([0, 1]) \leq 1$ . Therefore, using similar manipulations like those for bounding  $\Delta_1$ , we obtain the following.

$$\begin{aligned} \Delta_2 &= \mathbb{E}_{(p,y) \sim J}[(\hat{\kappa}(p) - y)\partial\ell(\kappa(p))] \\ &\leq \sup_{v \in [0,1]} \mathbb{E}_{(p,y) \sim J}[(y - \hat{\kappa}(p)) \text{sign}_+(\kappa(p) - v)] \end{aligned}$$

Since  $\mathcal{K}$  is closed under translations, there is  $\kappa' \in \mathcal{K}$  such that

$$\begin{aligned} \Delta_2 &\leq \mathbb{E}_{(p,y) \sim J}[(y - \hat{\kappa}(p)) \text{sign}_+(\kappa'(p) - 1/2)] \\ &\leq \sup_{c \in \mathcal{C}} \left| \mathbb{E}_{(p,y) \sim J}[(y - \hat{\kappa}(p)) c(p)] \right| \quad (\text{By the definition of } \mathcal{C} = \text{thr}(\mathcal{K})) \\ &\leq \varepsilon, \quad (\hat{\kappa} \in \text{calMA}_{\mathcal{C}}(J, \varepsilon)) \end{aligned}$$

which concludes the proof.  $\square$



The final ingredient of [Theorem 7.1](#) is the following result from [\[GHK<sup>+</sup>23\]](#) which reduces learning a predictor satisfying calibrated multiaccuracy to agnostic learning.

**Theorem 7.4** (Calibrated Multiaccuracy [\[GHK<sup>+</sup>23\]](#)). *Let  $\mathcal{C} \subseteq \{[0, 1] \rightarrow \{\pm 1\}\}$  and let  $\text{AL}$  be an  $(\varepsilon/3)$ -agnostic learner for  $\mathcal{C}$ . There is an algorithm that calls  $\text{AL}$   $O(\log(1/\delta)/\varepsilon^2)$  times, performs  $\text{poly}(1/\varepsilon) \log(1/\delta)$  additional operations, and outputs  $\hat{\kappa} \in \text{calMA}_{\mathcal{C}}(J, \varepsilon/2)$  with probability at least  $1 - \delta$ .*

## 7.2 Pool Adjacent Violators Is an Omnipredictor

Pool Adjacent Violators is a classical algorithm for recalibration [\[ABE<sup>+</sup>55\]](#), which finds a monotone post-processing of a predictor. It starts with a sample of  $\{(y_i, p_i)\}$  pairs. It starts from the Bayes optimal predictor  $\kappa(p_i) = \mathbb{E}[y_i | p_i]$ , and pools/merges any adjacent pair that violates monotonicity into a single interval  $I$  where the prediction is the conditional expectation  $\mathbb{E}[y | I]$ . We present the algorithm formally in [Algorithm 1](#).

---

### Algorithm 1: POOLADJACENTVIOLATORS( $S$ )

---

**Input:** Set  $S$  of  $m$  pairs of the form  $(p, y)$  where  $p \in [0, 1], y \in \{0, 1\}$

**Output:** A pair  $(\mathcal{O}, \mathcal{I})$  where  $\mathcal{O} = ((p_1, y_1), \dots, (p_m, y_m))$  is an ordering of the input set  $S$ , and  $\mathcal{I}$  is a sequence of disjoint intervals on  $[0, 1]$  that cover  $P = \{p_1, \dots, p_m\}$ .

- 1 Let  $\mathcal{O} = ((p_1, y_1), (p_2, y_2), \dots, (p_m, y_m))$  be such that  $p_1 \leq p_2 \leq \dots \leq p_m$ ;
  - 2 Let  $t = 0, \mathcal{I}^{(0)} = (I_1^{(0)}, \dots, I_m^{(0)})$ , where  $I_i^{(0)} = \{p_i\}$  for  $i \in [m]$ ;
  - 3 Set  $\bar{y}_i^{(0)} = y_i$  for all  $i \in [m]$ ;
  - 4 **while** there is  $j^* \in [m - t]$  such that  $\bar{y}_{j^*}^{(t)} \geq \bar{y}_{j^*+1}^{(t)}$  **do**
  - 5     Merge the sets  $I_{j^*}^{(t)}$  and  $I_{j^*+1}^{(t)}$ , i.e., set  $\mathcal{I}^{(t+1)} = (I_j^{(t+1)})_{j \in [m-t-1]}$ , where
 
$$I_j^{(t+1)} = \begin{cases} I_j^{(t)}, & \text{if } 0 \leq j < j^* \\ I_{j^*}^{(t)} \cup I_{j^*+1}^{(t)}, & \text{if } j = j^* \\ I_{j+1}^{(t)}, & \text{if } m - t - 1 \geq j > j^* \end{cases};$$
  - 6     Let  $\bar{y}_j^{(t+1)} = \frac{1}{|I_j^{(t+1)} \cap P|} \sum_{i: p_i \in I_j^{(t+1)}} y_i$  for all  $j \in [m - t]$ ;     ▷ Can be implemented in  $O(1)$
  - 7     Update  $t \leftarrow t + 1$ ;
  - 8 **Return**  $(\mathcal{O}, \mathcal{I}) = (\mathcal{O}, \mathcal{I}^{(t)})$ ;
- 

We show the following guarantee for PAV.

**Theorem 7.5** (Omniprediction through PAV). *For  $\varepsilon, \delta \in (0, 1)$ , Pool Adjacent Violators (PAV) run on  $O(\log(1/(\varepsilon\delta))/\varepsilon^2)$  samples is an algorithm that learns an  $(\varepsilon, \mathcal{M}_+)$ -omnipredictor with probability  $1 - \delta$  that runs in time  $\tilde{O}(1/\varepsilon^2)$ .*

In order to prove [Theorem 7.5](#), we first consider the empirical version of the omniprediction problem and show the following result.

**Theorem 7.6** (PAV guarantees). *Let  $S$  be a set of  $m$  pairs of the form  $(p, y)$  where  $p \in [0, 1]$  and  $y \in \{0, 1\}$ . There is an  $O(m \log m)$ -time algorithm ([Algorithm 1](#)) that computes a monotone post-processing  $\hat{\kappa} \in \mathcal{M}_+$  such that for any proper loss  $\ell \in \mathcal{L}^*$  we have:*

$$\sum_{(p, y) \in S} \ell(\hat{\kappa}(p), y) = \min_{\kappa \in \mathcal{M}_+} \sum_{(p, y) \in S} \ell(\kappa(p), y)$$

*Proof.* Let  $(\mathcal{O} = (p_i, y_i)_{i \in [m]}, \mathcal{I} = (I_j)_{j \in [m']})$  be the output of [Algorithm 1](#). The monotone post-processing  $\hat{\kappa}$  is obtained by setting for each  $i \in [m]$ :  $\hat{\kappa}(p_i) = \bar{y}_{I_j}$ , where  $j$  is such that  $p_i \in I_j$ , and interpolating linearly between the points  $p_i$ , in order to preserve monotonicity.

For any fixed proper loss  $\ell$ , we will show that  $\hat{\kappa}$  is an optimal monotone post-processing. To this end, we use an exchange argument to prove the correctness of the greedy approach of [Algorithm 1](#). In particular, we will show that at any step  $t \in \{0, 1, \dots, m-1\}$  of the algorithm, and for any  $j^* \in [m-t]$  such that  $\bar{y}_{j^*}^{(t)} \geq \bar{y}_{j^*+1}^{(t)}$ , there is an optimal monotone post-processing  $\tilde{\kappa}$  that gives the same value to all the points in the set  $I_{j^*}^{(t)} \cup I_{j^*+1}^{(t)}$ . Therefore, it is safe to merge  $I_{j^*}^{(t)}$  and  $I_{j^*+1}^{(t)}$ .

**Claim.** Assume that there is an optimal monotone post-processing that is constant on each of the intervals  $I_1, I_2, \dots, I_n$ . Let  $I_j, I_{j+1}$  be two intervals such that

$$\bar{y}_j = \mathbb{E}[y|p \in I_j] \geq \mathbb{E}[y|p \in I_{j+1}] = \bar{y}_{j+1}.$$

There exists an optimal post-processing  $\tilde{\kappa} \in \mathcal{M}_+$  such that  $\tilde{\kappa}(I_j) = \tilde{\kappa}(I_{j+1})$ .

*Proof.* Consider any post-processing function  $\kappa \in \mathcal{M}_+$  that is constant on each of the intervals  $I_1, \dots, I_n$ . Our goal is to find  $\tilde{\kappa} \in \mathcal{M}_+$  which does as well as  $\kappa$  for any proper loss  $\ell \in \mathcal{L}^*$ , and where  $\tilde{\kappa}(I_j) = \tilde{\kappa}(I_{j+1})$ . On the other intervals, we will have  $\tilde{\kappa} = \kappa$  and they both suffer the same loss, so we can ignore those intervals.

Let us write  $\kappa_i = \kappa(I_i)$ ,  $\tilde{\kappa}_i = \tilde{\kappa}(I_i)$  for  $i \in \{j, j+1\}$ . By monotonicity,  $\kappa_j \leq \kappa_{j+1}$ . If they are equal, we can take  $\tilde{\kappa} = \kappa$ , so assume the inequality is strict. In this case, we have  $\bar{y}_j \geq \bar{y}_{j+1}$  and  $\kappa_j < \kappa_{j+1}$ . We now consider three collectively exhaustive cases:

1.  $\kappa_j < \kappa_{j+1} \leq \bar{y}_j$ . We set  $\tilde{\kappa}_j = \tilde{\kappa}_{j+1} = \kappa_{j+1}$ . By [Lemma 3.2](#),

$$\ell(\tilde{\kappa}_j, \bar{y}_j) = \ell(\kappa_{j+1}, \bar{y}_j) \leq \ell(\kappa_j, \bar{y}_j)$$

so  $\tilde{\kappa}$  can only improve  $\kappa$  on interval  $I_j$ , while they agree on  $I_{j+1}$ .

2.  $\bar{y}_{j+1} \leq \kappa_j < \kappa_{j+1}$ . We set  $\tilde{\kappa}_j = \tilde{\kappa}_{j+1} = \kappa_j$ . By [Lemma 3.2](#),

$$\ell(\tilde{\kappa}_{j+1}, \bar{y}_{j+1}) = \ell(\kappa_j, \bar{y}_{j+1}) \leq \ell(\kappa_{j+1}, \bar{y}_{j+1})$$

so  $\tilde{\kappa}$  can only improve  $\kappa$  on interval  $I_{j+1}$ , while they agree on  $I_j$ .

3.  $\kappa_j < \bar{y}_{j+1} \leq \bar{y}_j < \kappa_{j+1}$ . Pick any  $a \in [\bar{y}_{j+1}, \bar{y}_j]$  and let  $\tilde{\kappa}_j = \tilde{\kappa}_{j+1} = a$ . Since  $\kappa_j < a \leq \bar{y}_{j+1}$ , by [Lemma 3.2](#),  $\ell(a, \bar{y}_j) \leq \ell(\kappa_j, \bar{y}_j)$ . Similarly, since  $\bar{y}_{j+1} \leq a < \kappa_{j+1}$ ,  $\ell(a, \bar{y}_{j+1}) \leq \ell(\kappa_{j+1}, \bar{y}_{j+1})$ . Hence  $\tilde{\kappa}$  can only improve  $\kappa$  on both intervals  $I_j, I_{j+1}$ .

This concludes the proof of the exchange argument.  $\square$

For a given loss  $\ell$ , let  $\kappa^*$  be an optimal monotone post-processing for the distribution  $S$ . Since  $\mathcal{I}^{(0)}$  is composed of singletons,  $\kappa^*$  is trivially constant on its intervals. The claim then implies that each iteration of PAV inductively preserves the property that there is an optimal monotone

postprocessing  $\kappa$  that is constant on intervals in  $\mathcal{I}^{(t)}$ . Considering a  $\kappa$  that is optimal for  $\mathcal{I}$ , and let  $\kappa_j$  be its value for  $p \in I_j$ . Denoting by  $\bar{y}_j$  the average of  $y$  on  $I_j$ , we have:

$$\begin{aligned} \mathbb{E}_{(p,y) \sim S}[\ell(\kappa(p), y)] &= \frac{1}{m} \sum_{j \in [m']} \sum_{i: p \in I_j} \ell(\kappa(p), y_i) \\ &= \frac{1}{m} \sum_{j \in [m']} \sum_{i: p \in I_j} \ell(\kappa_j, y_i) \\ &\geq \frac{1}{m} \sum_{j \in [m']} \sum_{i: p \in I_j} \ell(\bar{y}_j, y_i) \\ &= \mathbb{E}_{(p,y) \sim S}[\ell(\hat{\kappa}(p), y)], \end{aligned}$$

where the inequality follows from the fact that  $\ell$  is proper ([Definition 2.1](#)) and  $\bar{y}_j$  is the average of  $y$  on  $I_j$ . This implies that  $\hat{\kappa}$  is an optimal post-processing.

Finally, since the output of [Algorithm 1](#) satisfies  $\bar{y}_1 < \bar{y}_2 < \dots < \bar{y}_{m'}$ , it follows that  $\hat{\kappa}$  itself is monotone. The claim follows.  $\square$

*Remark 7.7.* Even though [Theorem 7.6](#) is stated for  $\mathcal{L}^*$ , it actually holds for all proper loss functions; in the case of PAV, the boundedness assumption is only needed for generalization.

We are now ready to prove [Theorem 7.5](#).

*Proof of Theorem 7.5.* It suffices to show that whenever  $|S| \geq C \log(1/\varepsilon\delta)/\varepsilon^2$  (for some sufficiently large constant  $C \geq 1$ , where  $S$  consists of i.i.d. samples from some distribution  $J$  over  $[0, 1] \times \{0, 1\}$ ), then, with probability at least  $1 - \delta$ , the following is true uniformly over all  $\kappa, \hat{\kappa} \in \mathcal{M}_+$ ,  $\ell \in \mathcal{L}^*$ :

$$\left| \mathbb{E}_{(p,y) \sim J}[\ell(\hat{\kappa}(p), y) - \ell(\kappa(p), y)] - \frac{1}{m} \sum_{(p,y) \in S} (\ell(\hat{\kappa}(p), y) - \ell(\kappa(p), y)) \right| \leq \varepsilon$$

The above inequality follows by combining the fact that  $\text{VCdim}(\text{thr}(\mathcal{M}_+)) = 1$  with [Lemma 4.2](#).  $\square$

### 7.3 Omniprediction Through Uniform-Mass Binning and Recalibration

We next analyze recalibration through uniform-mass binning. Binning is a long-established technique for measuring calibration [[Mil62](#), [San63](#)]. The method of uniform-mass binning was introduced by [[ZE01](#)] as the first binning-based approach not only for measuring calibration, but also for obtaining a calibrated predictor. We show that this natural algorithm yields omniprediction with respect to the class of generalized monotone post-processings. In other words, recalibration via uniform-mass binning improves the performance of the input predictor under every proper loss, with improvement at least as large as that achieved by the best generalized monotone post-processing. While previous work focused on the calibration properties of uniform-mass binning, we show that this method preserves the information encoded in the predictor—when measured by proper losses—at least as well as any generalized monotone post-processing.

We note that the omniprediction guarantee achieved by this method is stronger than the one we established for the PAV algorithm in [Theorem 7.5](#), since it applies to the larger class  $\mathcal{M}_r$ .

We define uniform-mass binning as follows.

**Definition 7.8** (Uniform-Mass Binning). Let  $P = (p_1, p_2, \dots, p_m)$  be a collection of  $m$  points in  $[0, 1]$ , with potential repetitions, and let  $\varepsilon \in (0, 1)$ . We say that a partition  $(I_j)_{j \in [t]}$  of  $[0, 1]$  is an  $\varepsilon$ -uniform-mass partition with respect to  $P$  if  $t \leq 2/\varepsilon$  and each  $I_j$  in the partition is an interval (open, closed, or half-open) for which at least one of the following holds:

- (Small buckets)  $|\{i \in [m] : p_i \in I_j\}| \leq \varepsilon m$ .
- (Overflow buckets) The interval  $I_j = [a, a]$  for some  $a \in [0, 1]$ .

Overflow buckets are necessary since it could be the case that  $p_i = a$  for say  $1/2$  the samples, in which case we create a separate bucket for it. We show that recalibration with uniform-mass binning achieves omniprediction with respect to the class of  $r$ -generalized monotone post-processings, with a number of buckets that is linear in  $r$ . While the resulting predictor  $\hat{\kappa}$  need not lie in  $\mathcal{M}_r$ , it is piecewise constant on  $r' = O(r/\varepsilon)$  intervals, hence it belongs to  $\mathcal{M}_{r'}$ .

**Theorem 7.9.** *Let  $\varepsilon \in (0, 1)$  and  $r \geq 1$ . Then, [Algorithm 2](#), run with parameter  $\varepsilon' = \varepsilon/8r$  on  $O(r^2 \log(1/\delta)/\varepsilon^4)$  samples learns an  $(\varepsilon, \mathcal{M}_r)$ -omnipredictor with probability  $1 - \delta$  and has time complexity  $O(r^2 \log(1/\delta)/\varepsilon^4)$ .*

---

**Algorithm 2:** UMB-RECALIBRATION( $S, \varepsilon'$ )

---

**Input:** Set  $S$  of  $m$  pairs of the form  $(p, y)$  where  $p \in [0, 1], y \in \{0, 1\}$ , parameter  $\varepsilon' \in (0, 1)$

**Output:** Post-processing function  $\hat{\kappa} : [0, 1] \rightarrow [0, 1]$

- 1 Create an  $\varepsilon'$ -uniform-mass partition  $(I_j)_{j \in [t]}$  w.r.t.  $(p_1, \dots, p_m)$  greedily, where  $t \leq 2/\varepsilon'$ ;
- 2 For all  $j \in [t]$ , let

$$\hat{\kappa}_j = \frac{\sum_{i \in [m]} y_i \mathbf{1}[p_i \in I_j]}{\sum_{i' \in [m]} \mathbf{1}[p_{i'} \in I_j]}.$$

- 3 Return the function  $\hat{\kappa}$  defined as

$$\hat{\kappa}(p) = \sum_{j=1}^t \hat{\kappa}_j \mathbf{1}[p \in I_j].$$


---

In order to prove [Theorem 7.9](#), we will use the following result regarding the sampling errors. The first inequality here is a standard uniform convergence bound. The second one captures the intuition that the per-bucket average is likely to be close to accurate for all buckets that are reasonably large. The somewhat unusual formalization of this condition that we use below better fits our application later, and allows for a tighter bound on the sample complexity.

**Lemma 7.10.** *Let  $\gamma \in (0, 1)$ , let  $S$  be a set of  $m$  i.i.d. samples from some distribution  $J$  over  $[0, 1] \times \{0, 1\}$ , and let  $(I_j, \hat{\kappa}_j)_{j \in [t]}$  be as defined in [Algorithm 2](#). If  $m \geq C \log(1/\delta)/\gamma^2$  for some sufficiently large constant  $C \geq 1$ , then the following hold with probability at least  $1 - \delta$ :*

$$\left| \mathbb{P}_J[p \in I_j] - \frac{1}{m} \sum_{i \in [m]} \mathbf{1}[p_i \in I_j] \right| \leq \gamma, \text{ for all } j \in [t].$$

$$|(\hat{\kappa}_j - \mathbb{E}_J[y|p \in I_j]) \cdot \Pr_J[p \in I_j]| \leq \gamma, \text{ for any } j \in [t];$$

*Proof.* The intervals  $(I_j)_{j \in [t]}$  in the output of the algorithm depend on the input examples  $(p_i, y_i)_{i \in [m]}$  which are drawn i.i.d. from some distribution  $J$ . Using standard uniform convergence arguments (as for proving [Lemma 4.2](#)), and the fact that the VC dimension of intervals over  $[0, 1]$  is 2, we have that with probability at least  $1 - \delta$ , the following guarantees hold as long as  $m \geq C \log(1/\delta)/\gamma^2$  for some sufficiently large constant  $C \geq 1$ .

$$\left| \frac{1}{m} \sum_{i \in [m]} y_i \mathbf{1}[p_i \in I_j] - \mathbb{E}_{(p,y) \sim J} [y \mathbf{1}[p \in I_j]] \right| \leq \frac{\gamma}{2}, \text{ for all } j \in [t] \quad (7.3)$$

$$\left| \frac{1}{m} \sum_{i \in [m]} \mathbf{1}[p_i \in I_j] - \mathbb{P}_{(p,y) \sim J} [p \in I_j] \right| \leq \frac{\gamma}{2}, \text{ for all } j \in [t]. \quad (7.4)$$

Fixing a  $j \in [t]$ , we write:

$$\begin{aligned} |(\hat{\kappa}_j - \mathbb{E}_J[y|p \in I_j]) \cdot \Pr_J[p \in I_j]| &= |\hat{\kappa}_j \cdot \Pr_J[p \in I_j] - \mathbb{E}_J[y \cdot \mathbf{1}[p \in I_j]]| \\ &= \left| \frac{\sum_{i \in [m]} y_i \mathbf{1}[p_i \in I_j]}{\sum_{i \in [m]} \mathbf{1}[p_i \in I_j]} \cdot \Pr_J[p \in I_j] - \mathbb{E}_J[y \cdot \mathbf{1}[p \in I_j]] \right| \\ &\leq \left| \frac{\sum_{i \in [m]} y_i \mathbf{1}[p_i \in I_j]}{\sum_{i \in [m]} \mathbf{1}[p_i \in I_j]} \cdot \Pr_J[p \in I_j] - \frac{1}{m} \sum_{i \in [m]} y_i \mathbf{1}[p_i \in I_j] \right| \\ &\quad + \left| \frac{1}{m} \sum_{i \in [m]} y_i \mathbf{1}[p_i \in I_j] - \mathbb{E}_J[y \cdot \mathbf{1}[p \in I_j]] \right| \\ &= \left| \frac{\sum_{i \in [m]} y_i \mathbf{1}[p_i \in I_j]}{\sum_{i \in [m]} \mathbf{1}[p_i \in I_j]} \right| \cdot \left| \Pr_J[p \in I_j] - \frac{1}{m} \sum_{i \in [m]} \mathbf{1}[p_i \in I_j] \right| \\ &\quad + \left| \frac{1}{m} \sum_{i \in [m]} y_i \mathbf{1}[p_i \in I_j] - \mathbb{E}_J[y \cdot \mathbf{1}[p \in I_j]] \right| \end{aligned}$$

Here in the first step, we have used the definition of  $\hat{\kappa}_j$ , and used triangle inequality in the second step. The second term in the final expression is easily seen to be at most  $\gamma/2$  by [Eq. \(7.3\)](#). Finally note that

$$\left| \frac{\sum_{i \in [m]} y_i \mathbf{1}[p_i \in I_j]}{\sum_{i \in [m]} \mathbf{1}[p_i \in I_j]} \right| \leq 1$$

for every instantiation of the samples, so that [Eq. \(7.4\)](#) implies a bound of  $\gamma/2$  on the first term as well. This concludes the proof of [Lemma 7.10](#).  $\square$

We are now ready to prove [Theorem 7.9](#). At a high level, we will split the decision loss for a loss  $\ell_v$  and a post-processing function  $\kappa$  across the buckets. We argue that except for a small number ( $O(r)$ ) of buckets, the sign of  $(\kappa(p) - v)$  is constant within the bucket, using the fact that  $\kappa \in \mathcal{M}_r$ . We separately handle these buckets using the fact that each such bucket (which cannot be an overflow bucket) has small probability mass. For the typical buckets that don't have a disagreement, the recalibration, if population-exact, would immediately ensure no decision loss. We will use [Lemma 7.10](#) to control the error that arises due to the potential inaccuracy of the sampling-based recalibration.

*Proof of Theorem 7.9.* With foresight, we set

$$\varepsilon' = \frac{\varepsilon}{8r}, \quad \gamma = \frac{\varepsilon\varepsilon'}{4} = \frac{\varepsilon^2}{32r}.$$

Our goal is to show the following inequality for all  $\kappa \in \mathcal{M}_r$  and  $\ell \in \mathcal{L}^*$ :

$$\mathbb{E}_{(p,y) \sim J}[\ell(\hat{\kappa}(p), y)] \leq \mathbb{E}_{(p,y) \sim J}[\ell(\kappa(p), y)] + \varepsilon \quad (7.5)$$

By Lemma 3.5, any proper loss  $\ell \in \mathcal{L}^*$  can be written as follows:

$$\ell(p, y) = \int_{[0,1]} \ell_v^+(p, y) d\mu_\ell^+(v) + \int_{[0,1]} \ell_v^-(p, y) d\mu_\ell^-(v),$$

where  $\mu_\ell$  is some measure over  $[0, 1]$  with  $\mu_\ell^+([0, 1]) + \mu_\ell^-([0, 1]) \leq 1$  and  $\ell_v^\pm(p, y) = -(y-v) \text{sign}_\pm(p-v)$ . Therefore, it suffices to show (7.5) for the losses  $(\ell_v^\pm)_{v \in [0,1]}$ . Below, we will focus on the losses  $(\ell_v^+)_{v \in [0,1]}$ , since the proof for  $(\ell_v^-)_{v \in [0,1]}$  will be analogous.

**Controlling Buckets of Disagreement.** We fix an arbitrary  $\kappa \in \mathcal{M}_r$  and  $v \in [0, 1]$ , and split  $[t]$  into two parts as follows. We let  $H = H_{\kappa,v} \subseteq [t]$  be as follows.

$$H = \left\{ j \in [t] : \exists p, q \in I_j \text{ such that } \text{sign}_+(\kappa(p) - v) \neq \text{sign}_+(\kappa(q) - v) \right\} \quad (7.6)$$

A key observation is that  $|H| \leq 2r$ , due to the definition of  $\mathcal{M}_r$  (Definition 2.2). In particular, since  $\{p : \kappa(p) \geq v\}$  can be expressed as a union of at most  $r$  disjoint intervals  $\mathcal{I}$ , the number of sign changes of  $\kappa(p) - v$  as  $p$  increases from 0 to 1 is at most  $2r$  (at the endpoints of the intervals). Moreover, any  $j \in H$  corresponds to at least one distinct point of sign change for  $\kappa(p) - v$ , and, hence,  $|H| \leq 2r$ .

Note that whenever  $|I_j| = 1$ , we have  $j \notin H$ . Recall that, due to the construction of  $(I_j)_{j \in [t]}$  (Algorithm 2), for any  $j$  such that  $|I_j| > 1$  we have  $|I_j \cap \{p_i : i \in [m]\}| \leq \varepsilon' m$ . Due to the uniform convergence bound of Lemma 7.10 we overall have:

$$\mathbb{P}_J[p \in I_j] \leq \varepsilon' + \gamma \text{ for all } j \in H \quad (7.7)$$

**Handling typical buckets.** For an interval  $I_j$  not in  $H$ , the  $\text{sign}_+(\kappa(p) - v)$  is constant throughout the interval, and the same is true for  $\hat{\kappa}$ , by construction. Let  $s_j = \text{sign}_+(\kappa(p) - v)$  for  $p \in I_j$ ,  $j \in [t] \setminus H$  and similarly  $\hat{s}_j = \text{sign}_+(\hat{\kappa}(p) - v)$ . For such a bucket, we can write

$$\begin{aligned} & \mathbb{E}[\ell_v^+(\hat{\kappa}(p), y) \cdot \mathbf{1}[p \in I_j]] - \mathbb{E}[\ell_v^+(\kappa(p), y) \cdot \mathbf{1}[p \in I_j]] \\ &= \mathbb{E}[-(\mathbb{E}[y|\mathbf{1}[p \in I_j]] - v) \cdot (\hat{s}_j - s_j) \cdot \mathbf{1}[p \in I_j]] \\ &= \mathbb{E}[-(\hat{\kappa}_j - v) \cdot (\hat{s}_j - s_j) \cdot \mathbf{1}[p \in I_j]] + \mathbb{E}[-(\mathbb{E}[y|\mathbf{1}[p \in I_j]] - \hat{\kappa}_j) \cdot (\hat{s}_j - s_j) \cdot \mathbf{1}[p \in I_j]]. \end{aligned}$$

Observe that  $-(\hat{\kappa}_j - v)\hat{s}_j = -|\hat{\kappa}_j - v| \leq -(\hat{\kappa}_j - v)s_j$  so that the first term is bounded above by zero. On the other hand, the second term is controlled by Lemma 7.10:

$$\begin{aligned} \mathbb{E}[-(\mathbb{E}[y|\mathbf{1}[p \in I_j]] - \hat{\kappa}_j) \cdot (\hat{s}_j - s_j) \cdot \mathbf{1}[p \in I_j]] &\leq 2 \cdot \mathbb{E}[(\mathbb{E}[y|\mathbf{1}[p \in I_j]] - \hat{\kappa}_j) \cdot \mathbf{1}[p \in I_j]] \\ &\leq 2\gamma. \end{aligned}$$

Thus we have shown that for any interval  $I_j \notin H$ :

$$\mathbb{E}[\ell_v^+(\hat{\kappa}(p), y) \cdot \mathbf{1}[p \in I_j]] - \mathbb{E}[\ell_v^+(\kappa(p), y) \cdot \mathbf{1}[p \in I_j]] \leq 2\gamma \quad (7.8)$$

**Putting it Together.** We are now ready to prove the theorem. We will split the buckets into the buckets of disagreement and the rest, and use the bounds above to control the total error. We write

$$\begin{aligned}
& \mathbb{E}_{(p,y) \sim J}[\ell(\widehat{\kappa}(p), y)] - \mathbb{E}_{(p,y) \sim J}[\ell(\kappa(p), y)] \\
&= \sum_{j \in [t]} \mathbb{E}_{(p,y) \sim J}[(\ell(\widehat{\kappa}(p), y) - \ell(\kappa(p), y)) \cdot \mathbf{1}[p \in I_j]] \\
&= \sum_{j \in H} \mathbb{E}_{(p,y) \sim J}[(\ell(\widehat{\kappa}(p), y) - \ell(\kappa(p), y)) \cdot \mathbf{1}[p \in I_j]] \\
&\quad + \sum_{j \in [t] \setminus H} \mathbb{E}_{(p,y) \sim J}[(\ell(\widehat{\kappa}(p), y) - \ell(\kappa(p), y)) \cdot \mathbf{1}[p \in I_j]] \\
&\leq 2r \cdot (\varepsilon' + \gamma) + (2/\varepsilon') \cdot (2\gamma) \\
&\leq 4r\varepsilon' + 4\gamma/\varepsilon' \\
&\leq \varepsilon.
\end{aligned}$$

Here we have bounded the sum over  $H$  and  $[t] \setminus H$  using [Eq. \(7.7\)](#) and [Eq. \(7.8\)](#) respectively. The claim follows.  $\square$

## References

- [ABE<sup>+</sup>55] Miriam Ayer, H Daniel Brunk, George M Ewing, William T Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, pages 641–647, 1955. 3, 4, 11, 29, 31
- [BdP13] Niko Brümmer and Johan A. du Preez. The PAV algorithm optimizes binary proper scoring rules. *ArXiv*, abs/1304.2331, 2013. 11, 12
- [BGHN23a] 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. 1, 2, 4, 8, 19
- [BGHN23b] Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? In *Advances in Neural Information Processing Systems*, volume 36, pages 72071–72095, 2023. 1, 2, 4, 8, 19
- [BHJB25] Eugène Berta, David Holzmüller, Michael I. Jordan, and Francis Bach. Rethinking early stopping: Refine, then calibrate, 2025. 3
- [BN24] Jaroslaw Blasiok and Preetum Nakkiran. Smooth ECE: principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. 1, 2
- [CDV24] Sílvia Casacuberta, Cynthia Dwork, and Salil Vadhan. Complexity-theoretic implications of multicalibration. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024*, page 1071–1082, 2024. 1
- [Daw85] A Philip Dawid. Calibration-based empirical probability. *The Annals of Statistics*, pages 1251–1274, 1985. 1
- [DKR<sup>+</sup>21] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *ACM Symposium on Theory of Computing (STOC’21)*, 2021. 1, 2, 20
- [Fel09] Vitaly Feldman. Distribution-specific agnostic boosting. *arXiv preprint arXiv:0909.2927*, 2009. 7
- [FV98] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998. 1, 3
- [GH25] Parikshit Gopalan and Lunjia Hu. Calibration through the lens of indistinguishability. *ACM SIGecom Exchanges*, 23(1), July 2025. 2, 8
- [GHK<sup>+</sup>23] Parikshit Gopalan, Lunjia Hu, Michael P Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, 2023. 2, 9, 10, 11, 19, 20, 29, 31
- [GHR24] Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. On computationally efficient multi-class calibration. In *The Thirty Seventh Annual Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1983–2026. PMLR, 2024. 2, 13, 16, 17



- [GKR<sup>+</sup>22] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Ominipredictors. In *Innovations in Theoretical Computer Science (ITCS'2022)*, 2022. [1](#), [10](#)
- [GKSZ22] Parikshit Gopalan, Michael P. Kim, Mihir Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 3193–3234. PMLR, 2022. [2](#), [8](#)
- [GPSW17] 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. [3](#), [12](#)
- [GR21] Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International conference on machine learning*, pages 3942–3952. PMLR, 2021. [3](#), [4](#), [11](#), [12](#), [29](#)
- [HJTY24] Lunjia Hu, Arun Jambulapati, Kevin Tian, and Chutong Yang. Testing calibration in nearly-linear time. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [2](#)
- [HKRR18] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML, 2018*. [1](#)
- [HSLW23] Jason D. Hartline, Liren Shan, Yingkai Li, and Yifan Wu. Optimal scoring rules for multi-dimensional effort. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 2624–2650. PMLR, 2023. [5](#)
- [HV25] Lunjia Hu and Salil P. Vadhan. Generalized and unified equivalences between hardness and pseudoentropy. *CoRR*, abs/2507.05972, 2025. [1](#)
- [HW24] 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. [1](#), [3](#), [4](#), [6](#), [7](#), [8](#), [9](#), [14](#), [45](#)
- [HWY25] Jason D. Hartline, Yifan Wu, and Yunran Yang. Smooth calibration and decision making. In *11th Symposium on Foundations of Responsible Computing (FORC 2025)*, volume 329 of *LIPICs*, pages 16:1–16:22. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. [11](#)
- [KF08] Sham Kakade and Dean Foster. Deterministic calibration and Nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008. [1](#), [2](#), [19](#)
- [KK09] Adam Kalai and Varun Kanade. Potential-based agnostic boosting. In *Advances in Neural Information Processing Systems*, volume 22, 2009. [7](#)
- [KLST23] 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. [3](#), [4](#), [5](#), [8](#), [9](#), [14](#)

- [KSS94] Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. *Mach. Learn.*, 17(2-3):115–141, 1994. 7, 46
- [Mil62] Robert G Miller. Statistical prediction by discriminant analysis. In *Statistical prediction by discriminant analysis*, pages 1–54. Springer, 1962. 12, 33
- [MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. 43
- [NC05] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In Luc De Raedt and Stefan Wrobel, editors, *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005)*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM, 2005. 3
- [NCH15] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015. 12
- [OKK25] Princewill Okoroafor, Robert Kleinberg, and Michael P. Kim. Near-optimal algorithms for omniprediction. *CoRR*, abs/2501.17205, 2025. 1, 2, 20
- [Pla00] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000. 3, 4
- [RCSM22] 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. 12
- [RSB<sup>+</sup>25] Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. Can a calibration metric be both testable and actionable? In *The Thirty Eighth Annual Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4937–4972. PMLR, 2025. 1, 2, 3, 4, 8, 9, 20
- [San63] Frederick Sanders. On subjective probability forecasting. *Journal of Applied Meteorology and Climatology*, 2(2):191–201, 1963. 12, 33
- [SSH23] Zeyu Sun, Dogyoon Song, and Alfred Hero. Minimum-risk recalibration of classifiers. *Advances in Neural Information Processing Systems*, 36:69505–69531, 2023. 3, 4, 11, 12, 29
- [ZE01] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer, 2001. 3, 4, 11, 12, 29, 33
- [ZE02] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multi-class probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699. ACM, 2002. 3, 4, 11
- [ZKS<sup>+</sup>21] Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 22313–22324, 2021. 1

## A Why Generalized Monotone Functions

Here we discuss why generalized monotone functions are a natural class of post-processing functions. Intuitively, monotonicity is a natural constraint on post-processing functions to apply to predictors if we believe that the predictors are reasonably good to begin with: if  $p = 0.7$ , that ought to mean that the probability of the label 1 is higher than if  $p = 0.3$ . However, this might not be true for two close-by values like 0.7 and 0.6999. Thus it is natural to relax the condition and allow some violations of monotonicity (see [Figure 1](#) for an example).

A first attempt might be through allowing small total variation. An increasing sequence  $I_n$  of length  $n$  in  $[0, 1]$  is  $(p_i)_{i \in [n]}$  such that  $0 \leq p_1 < p_2 < \dots < p_n \leq 1$ . For a function  $\kappa : [0, 1] \rightarrow [0, 1]$ , we define its total variation as

$$\text{tv}(\kappa) = \lim_{n \rightarrow \infty} \sup_{I_n} \sum_{i=1}^{n-1} |\kappa(p_{i+1}) - \kappa(p_i)|$$

where the supremum is over all increasing sequences  $I_n$  of length  $n$ . The class  $\text{tv}(f) \leq 1$  contains both monotone and Lipschitz functions. But we have seen that  $\text{VCdim}(\text{thr}(\text{Lip})) = \infty$  whereas  $\text{VCdim}(\text{thr}(\mathcal{M}_+)) = 1$ , so they are very different from CDL viewpoint.

Is there a strengthening of  $\text{tv}$  that rules out arbitrary Lipschitz function but extends monotone functions? We show that  $\mathcal{M}_r$  is such a class.

**Definition A.1.** For  $v \in (0, 1)$ , let the *crossing number* at  $v$  denoted  $\text{cr}_v(\kappa)$  denote the largest  $n$  for which there is a strictly increasing sequence  $(p_i)_{i \in [n+1]}$  such that  $\text{sign}(\kappa(p_i) - v)$  alternates. Let

$$\begin{aligned} \text{cr}(\kappa) &= \sup_{v \in [0,1]} \text{cr}_v(\kappa), \\ \text{cr}(\mathcal{K}) &= \sup_{\kappa \in \mathcal{K}} \text{cr}(\kappa) \end{aligned}$$

It is easy to show that  $\text{cr}(f)$  enjoys the following properties:

- $\text{cr}(\kappa)$  is within a constant factor of the smallest  $r$  for which  $\kappa \in \mathcal{M}_r$ .
- We have  $\text{VCdim}(\text{thr}(\mathcal{K})) \leq \text{cr}(\mathcal{K})$ , since the alternating sequence of signs on  $\text{cr}(\mathcal{K}) + 1$  points is not realizable within  $\text{thr}(\mathcal{K})$ .
- $\text{tv}(\kappa) \leq \text{cr}(\kappa)$ . We can see  $\text{tv}(\kappa)$  as a bound on  $\mathbb{E}_v[\text{cr}_v(\kappa)]$  over uniformly random  $v \in [0, 1]$ , whereas  $\text{cr}(\kappa)$  bounds  $\text{cr}_v(\kappa)$  in the worst case. For more detail, see the following lemma.

**Lemma A.2.**  $\text{tv}(\kappa) \leq \text{cr}(\kappa)$  for all  $\kappa : [0, 1] \rightarrow [0, 1]$ .

*Proof.* Fix any increasing sequence  $I_n$  of points  $0 \leq p_1 < p_2 < \dots < p_n \leq 1$ . For any particular index  $2 \leq i \leq n$ , the probability over a uniformly random value  $v \sim [0, 1]$  that  $\text{sign}(\kappa(p_{i-1}) - v) \neq \text{sign}(\kappa(p_i) - v)$  is precisely  $|\kappa(p_i) - \kappa(p_{i-1})|$ . Let  $\text{cr}_v(\kappa, p)$  denote the number of consecutive indices for which this sign change occurs. Then, by linearity of expectation,

$$\mathbb{E}_v[\text{cr}_v(\kappa, p)] = \sum_{i=2}^n \Pr_v[\text{sign}(\kappa(p_{i-1}) - v) \neq \text{sign}(\kappa(p_i) - v)] = \sum_{i=2}^n |\kappa(p_{i-1}) - \kappa(p_i)|.$$

The supremum of the right side over all sequences  $I_n$  is precisely  $\text{tv}(\kappa)$ , by definition. The supremum of the left hand side over all sequences  $I_n$  is

$$\sup_{I_n} \mathbb{E}_v[\text{cr}_v(\kappa, p)] \leq \sup_v \sup_{I_n} \text{cr}_v(\kappa, p) = \sup_v \text{cr}_v(\kappa) = \text{cr}(\kappa).$$

□

Thus  $\mathcal{M}_r = \{\kappa : \text{cr}(\kappa) \leq O(r)\}$  corresponds to a class of functions that relaxes monotonicity, but excludes functions whose thresholds have high VC dimension, like Lipschitz functions.

## B Additional Proofs

### B.1 Additional Proofs From Section 3

*Proof of Lemma 3.2.* Take  $\varphi, \varphi'$  as in the characterization in Lemma 3.1, so that

$$\ell(p, q) = \varphi(p) + \varphi'(p)(q - p).$$

Then we have

$$\begin{aligned} \ell(a, c) - \ell(b, c) &= \varphi(a) + \varphi'(a)(c - a) - (\varphi(b) + \varphi'(b)(c - b)) \\ &= \underbrace{\varphi(a) + \varphi'(a)(b - a) - \varphi(b)}_{(*)} + \underbrace{(\varphi'(a) - \varphi'(b))(c - b)}_{(**}), \end{aligned}$$

where nonnegativity of  $(*)$  and  $(**)$  follow from concavity of  $\varphi$ . The other inequality is analogous.  $\square$

*Proof of Lemma 3.3.* By Lemma 3.2, the function  $\ell(p, 0)$  is nondecreasing in  $p$ , and the function  $\ell(p, 1)$  is nonincreasing in  $p$ . We also have  $|\ell(p, 1) - \ell(p, 0)| \leq 1$  for any  $\ell \in \mathcal{L}^*$ . Thus, for any  $0 \leq p \leq q \leq 1$ ,

$$\ell(p, 0) \leq \ell(q, 0) \leq \ell(q, 1) + 1.$$

Considering more cases like this, we see that  $|\ell(p, 0) - \ell(q, 1)| \leq 1$  for all  $p, q \in [0, 1]$ . Combining this inequality with  $|\ell(p, 1) - \ell(p, 0)| \leq 1$ , we see that any two outputs of  $\ell$  on distinct inputs differ by at most 2. Equivalently, we conclude that the range of any  $\ell \in \mathcal{L}^*$  is contained in an interval of length at most 2.  $\square$

*Proof of Corollary 3.6.* First, by Lemma 3.5, we know that any loss  $\ell \in \mathcal{L}^*$  that is not of the form  $\ell_v^+$  or  $\ell_v^-$  or a constant function is redundant and need not be considered in the supremum. Next, for any fixed  $(p, y) \in [0, 1] \times \{0, 1\}$  and  $v \in [0, 1]$ , we have  $\lim_{w \rightarrow v^+} \ell_w^+(p, y) = \ell_v^+(p, y)$ . Therefore, by the bounded convergence theorem, the losses  $\ell_v^-$  for  $v \in [0, 1]$  are redundant if we take the supremum over all  $\ell_w^+$  for  $w \in [0, 1]$ . Note that  $\ell_1^-(p, y) = y - 1$  is also redundant because it does not depend on  $p$ . Similarly, if  $v \in (0, 1]$ , then  $\lim_{w \rightarrow v^-} \ell_w^-(p, y) = \ell_v^+(p, y)$ . Therefore, the losses  $\ell_v^+$  for  $v \in (0, 1]$  are redundant if we take the supremum over  $\ell_w^-$  losses for  $w \in [0, 1]$ . Note that  $\ell_0^+(p, y) = -y$  is also redundant because it does not depend on  $p$ . The same is true of constant functions.  $\square$

### B.2 Additional Proofs From Section 4

*Proof of Lemma 4.2.* We define the following quantities for  $v \in [0, 1]$  and  $\kappa \in \mathcal{K}$ .

$$\begin{aligned} Q_{v,\kappa}^+ &= \mathbb{E}_{(p,y) \sim J} [\ell_v^+(p, y) - \ell_v^+(\kappa(p), y)] \text{ and } Q_{v,\kappa}^- = \mathbb{E}_{(p,y) \sim J} [\ell_v^-(p, y) - \ell_v^-(\kappa(p), y)] \\ \widehat{Q}_{v,\kappa}^+ &= \mathbb{E}_{(p,y) \sim S} [\ell_v^+(p, y) - \ell_v^+(\kappa(p), y)] \text{ and } \widehat{Q}_{v,\kappa}^- = \mathbb{E}_{(p,y) \sim S} [\ell_v^-(p, y) - \ell_v^-(\kappa(p), y)] \end{aligned}$$

Due to Corollary 3.6, it suffices to show that, with probability at least  $1 - \delta$  over the choice of  $S$ , we have the following:

$$|Q_{v,\kappa}^\star - \widehat{Q}_{v,\kappa}^\star| \leq \varepsilon, \text{ for all } v \in [0, 1], \kappa \in \mathcal{K}, \star \in \{+, -\} \quad (\text{B.1})$$

For the remainder of the proof, we will focus on  $|Q_{v,\kappa}^+ - \widehat{Q}_{v,\kappa}^+|$ , since the other case follows identically, with the additional observation that, for translation-invariant classes  $\mathcal{K}$ , we have  $\text{VCdim}(\text{thr}(\mathcal{K})) = \text{VCdim}(\text{thr}_-(\mathcal{K}))$ , where  $\text{thr}_-(\mathcal{K}) = \{p \mapsto \text{sign}_-(\kappa(p) - 1/2) : \kappa \in \mathcal{K}\}$ .

We will show that condition (B.1) holds with probability at least  $1 - \delta$  as long as the size of  $S$  is  $m \geq \frac{Cd}{\varepsilon^2} \log(\frac{1}{\varepsilon\delta})$  for some large enough constant  $C \geq 1$ . To this end, recall that  $\ell_v^+(p, y) = -(y - v) \text{sign}_+(p - v)$ . Therefore:

$$\begin{aligned} Q_{v,\kappa}^+ &= \mathbb{E}_{(p,y) \sim J} [(y - v)(\text{sign}_+(\kappa(p) - v) - \text{sign}_+(p - v))] \\ \widehat{Q}_{v,\kappa}^+ &= \mathbb{E}_{(p,y) \sim S} [(y - v)(\text{sign}_+(\kappa(p) - v) - \text{sign}_+(p - v))] \end{aligned}$$

For the following, we use  $\text{sign}$  to denote  $\text{sign}_+$  and  $Q_{v,\kappa}, \widehat{Q}_{v,\kappa}$  to denote  $Q_{v,\kappa}^+, \widehat{Q}_{v,\kappa}^+$ . Since  $\mathcal{K}$  is translation invariant, for any  $\kappa \in \mathcal{K}$  and  $v \in [0, 1]$ , there is  $\kappa' \in \mathcal{K}$  such that  $\kappa'(p) = [\kappa(p) + \frac{1}{2} - v]_0^1$ . Then, we have  $\text{sign}(\kappa(p) - v) = \text{sign}(\kappa(p) + 1/2 - v - 1/2) = \text{sign}(\kappa'(p) - 1/2)$ , for all  $p \in [0, 1]$ , and, therefore:

$$\begin{aligned} Q_{v,\kappa} &= \mathbb{E}_{(p,y) \sim J} [(y - v)(\text{sign}(\kappa'(p) - 1/2) - \text{sign}(p - v))] \\ \widehat{Q}_{v,\kappa} &= \mathbb{E}_{(p,y) \sim S} [(y - v)(\text{sign}(\kappa'(p) - 1/2) - \text{sign}(p - v))] \end{aligned}$$

Let  $V = \{\frac{i\varepsilon}{16} : i = 0, 1, \dots, \lfloor \frac{16}{\varepsilon} \rfloor\} \cup \{1\}$ , and  $\pi_\varepsilon(v) = \arg \min_{v' \in V} |v - v'|$ . We define the following quantities.

$$\begin{aligned} Q_{v,\kappa}^V &= \mathbb{E}_{(p,y) \sim J} [(y - \pi_\varepsilon(v))(\text{sign}(\kappa'(p) - 1/2) - \text{sign}(p - v))] \\ \widehat{Q}_{v,\kappa}^V &= \mathbb{E}_{(p,y) \sim S} [(y - \pi_\varepsilon(v))(\text{sign}(\kappa'(p) - 1/2) - \text{sign}(p - v))] \end{aligned}$$

Note that, by the choice of  $V$ , we have  $|Q_{v,\kappa}^V - Q_{v,\kappa}| \leq \varepsilon/8$ , and, similarly,  $|\widehat{Q}_{v,\kappa}^V - \widehat{Q}_{v,\kappa}| \leq \varepsilon/8$ , for all  $v \in [0, 1]$  and  $\kappa \in \mathcal{K}$ . Therefore, it suffices to show that with probability at least  $1 - \delta$ , the following holds for any  $v \in [0, 1]$ , any  $v' \in V$ , and any  $\kappa' \in \mathcal{K}$ :

$$\left| \mathbb{E}_{(p,y) \sim J} [(y - v')(\text{sign}(\kappa'(p) - 1/2) - \text{sign}(p - v))] - \mathbb{E}_{(p,y) \sim S} [(y - v')(\text{sign}(\kappa'(p) - 1/2) - \text{sign}(p - v))] \right| \leq \frac{\varepsilon}{4}.$$

Since the size of  $V$  is  $O(1/\varepsilon)$ , it suffices to prove that the above bound holds for each individual  $v' \in V$  (but uniformly over  $v \in [0, 1]$  and  $\kappa \in \mathcal{K}$ ) with probability  $1 - O(\varepsilon\delta)$ . The desired result would then follow by a union bound. For each  $v' \in V$ , the desired bound is true due to standard uniform convergence arguments, combined with the fact that the VC dimension of the class  $\{p \mapsto \text{sign}(\kappa(p) - 1/2) : \kappa \in \mathcal{K}\}$  is  $d$ , and the VC dimension of the class  $\{p \mapsto \text{sign}(p - v) : v \in [0, 1]\}$  is 1. In particular, we may combine the following results from [MRT18]: Corollary 3.8 (which gives a bound on the Rademacher complexity of a binary class in terms of the associated growth function), Theorem 3.17 (Sauer's Lemma, which bounds the growth function in terms of the VC dimension), and Theorem 11.3 (which gives a generalization bound for regression with respect to bounded and Lipschitz losses, in terms of the Rademacher complexity of the underlying function class). The choice of  $O(\varepsilon\delta)$  for the failure probability only incurs an additive term of  $O(\log(1/\varepsilon))$ , and, therefore, our choice for the number of samples  $m$  suffices to achieve the desired result.  $\square$

*Proof of Theorem 4.1 Upper Bound.* Let  $\mathcal{A}_1$  be the algorithm that receives a set  $S$  of  $m$  i.i.d. examples from some unknown distribution  $J$  over  $[0, 1] \times \{0, 1\}$ , where  $m \geq C \cdot d \log(1/\varepsilon)/\varepsilon^2$ , for some sufficiently large universal constant  $C \geq 1$ , and does the following:

1. Compute the following quantity:

$$\text{CDL}_{\mathcal{K}}(S) = \sup_{\kappa \in \mathcal{K}, \ell \in \mathcal{L}^*} \frac{1}{m} \sum_{(p,y) \in S} (\ell(p,y) - \ell(\kappa(p), y))$$

2. If  $\text{CDL}_{\mathcal{K}}(S) \leq \alpha - \epsilon/2$ , then output **Accept**. Otherwise, output **Reject**.

By [Lemma 4.2](#), with probability at least  $2/3$ , we have  $|\text{CDL}_{\mathcal{K}}(J) - \text{CDL}_{\mathcal{K}}(S)| \leq \epsilon/2$ . Under this event, if  $\text{CDL}_{\mathcal{K}}(J) \leq \alpha - \epsilon$ , then algorithm  $\mathcal{A}_1$  will accept. On the other hand, if  $\text{CDL}_{\mathcal{K}}(J) > \alpha$ , then  $\mathcal{A}_1$  will reject.  $\square$

*Proof of Corollary 4.7.* The proof of the upper bound is essentially the same as in [Theorem 4.1](#). The key ingredient of that proof was the uniform convergence bound stated in [Lemma 4.2](#), which holds for a supremum over all  $\ell \in \mathcal{L}^*$ . In particular, it continues to hold if we only take the supremum over  $\ell \in \mathcal{L}_{\mu\text{-sc}}^*$ , so the proof still goes through. For the lower bound, we can no longer rely on our argument based on the  $\ell_{1/2}^+$  loss, since its associated function  $\varphi_{1/2}(p) = -|p - 1/2|$  is not strongly concave. To circumvent this issue, consider the following version of the squared loss:

$$\ell_{\text{sq}}(p, y) = (y - p)^2.$$

Then  $\partial \ell_{\text{sq}}(p) = 1 - 2p \in [-1, +1]$ , so the function  $\ell_{\text{sq}}$  indeed belongs to the class  $\mathcal{L}^*$ . Moreover,  $\varphi_{\text{sq}}(p) = \ell_{\text{sq}}(p, p) = p(1 - p)$  is 2-strongly concave, so  $\ell_{\text{sq}} \in \mathcal{L}_{2\text{-sc}}^*$ . Next, given any  $\ell \in \mathcal{L}^*$  and  $\mu > 0$ , we define the following convex combination, which belongs to  $\mathcal{L}_{\mu\text{-sc}}^*$ :

$$\ell_{\mu} = \frac{\mu}{2} \ell_{\text{sq}} + \left(1 - \frac{\mu}{2}\right) \ell.$$

We will use  $\ell_{\mu}$  to study the effect of restricting to  $\mu$ -strongly convex proper losses. First, since we are considering a restricted class of loss functions, we clearly have  $\text{CDL}_{\mathcal{L}_{\mu\text{-sc}}^*, \mathcal{K}} \leq \text{CDL}_{\mathcal{K}}$ . Conversely, let  $\kappa$  be a post-processing function that improves  $\ell$  by at least  $\alpha$ , meaning that

$$\mathbb{E}[\ell_{\text{sq}}(\kappa(p), y)] \leq \mathbb{E}[\ell_{\text{sq}}(p, y)] - \alpha.$$

Since the convex combination  $\ell_{\mu}$  puts  $1 - \mu/2$  weight on  $\ell$ , the post-processing  $\kappa$  must improve the loss on this part of  $\ell_{\mu}$  by at least  $(1 - \mu/2)\alpha$ . Although  $\kappa$  may worsen the loss on  $\ell_{\text{sq}}$  arbitrarily, the convex combination  $\ell_{\mu}$  puts only  $\mu/2$  weight on  $\ell_{\text{sq}}$ , so it must worsen the loss on this part of  $\ell_{\mu}$  by at most  $\mu_2 \cdot 2$  (recall that all losses in  $\mathcal{L}^*$ , such as  $\ell_{\text{sq}}$  have range bounded in an interval of length 2, by [Lemma 3.3](#)). In total,  $\kappa$  must improve the loss on  $\ell_{\mu}$  by at least

$$\left(1 - \frac{\mu}{2}\right)\alpha - \mu \geq \alpha - 2\mu.$$

It follows immediately that an  $(\alpha, \beta)$ -auditor for  $\text{CDL}_{\mathcal{K}}$  is implied by an  $(\alpha - 2\mu, \beta)$ -auditor for  $\text{CDL}_{\mathcal{L}_{\mu\text{-sc}}^*, \mathcal{K}}$ . Thus, setting  $\alpha = 1/8$  and  $\beta = 0$ , our lower bound for auditing carries over to the case of  $\mathcal{L}_{\mu\text{-sc}}^*$ , as claimed.  $\square$

*Proof of Theorem 4.9.* We will prove the upper and lower bounds separately.

**Upper Bound.** From the loss OI lemma ([Lemma 5.4](#)), we have that  $\ell(p, y) - \ell(\kappa(p), y) \leq (\partial \ell(\kappa(p)) - \partial \ell(p))(y - p)$ . The function  $w'(p) = \partial \ell(\kappa(p)) - \partial \ell(p)$  is 6-Lipschitz, because  $\kappa$  is 2-Lipschitz and  $\partial \ell$  is 2-Lipschitz. Therefore, we have

$$\text{smCDL}(J) \leq \sup_{w': 6\text{-Lipschitz}} \mathbb{E}[w'(p)(y - p)] \leq 6 \cdot \text{smCE}(J)$$

**Lower Bound.** Suppose that there is a 1-Lipschitz function  $w : [0, 1] \rightarrow [-1, 1]$  such that

$$\text{smCE}(J) = \mathbb{E}_{(p,y) \sim J}[(y - p) \cdot w(p)] = \alpha.$$

Then, the post-processing function  $\kappa(p) = [p + \alpha w(p)]_0^1$  is 2-Lipschitz and satisfies:

$$\mathbb{E}_{(p,y) \sim J}[(y - \kappa(p))^2] \leq \mathbb{E}_{(p,y) \sim J}[(y - p)^2] - \alpha^2$$

The squared loss  $\ell_{\text{sq}}(p, y) = (p - y)^2/2$  is 1-Lipschitz and proper. Therefore,  $\text{smCDL}(J) \geq \alpha^2/2$ .  $\square$

## C Tightness of the Weight-Restricted Calibration Characterization

In this section, we show that the quadratic gap in [Theorem 5.2](#) is essentially tight, using the example of the class of monotonically nondecreasing post-processings  $\mathcal{K} = \mathcal{M}_+$ .

**Theorem C.1.** *There exist distributions  $J_1, J_2$  over pairs  $(p, y) \in [0, 1] \times \{0, 1\}$  such that*

$$\text{CDL}_{\mathcal{M}_+}(J_1) \gtrsim \text{CE}_{\text{Int}}(J_1)$$

and

$$\text{CDL}_{\mathcal{M}_+}(J_2) \lesssim \text{CE}_{\text{Int}}(J_2)^2.$$

*Proof.* Our examples are essentially the same ones used by [\[HW24\]](#) to establish the tightness of their relationship between  $\text{CDL}_{\mathcal{K}^*}$  and ECE, which also has a quadratic gap.

For  $J_1$ , suppose that  $p = 1 - \varepsilon$  and  $y = 1$  deterministically. Then,

$$\text{CE}_{\text{Int}}(J_1) \leq \text{ECE}(J_1) = \varepsilon,$$

but if we set  $v = 1 - \varepsilon/2$  and consider the monotonic post-processing  $\kappa(p) = p + \varepsilon$ , then

$$\text{CDL}_{\mathcal{M}_+}(J_1) \geq \mathbb{E}[\ell_v^+(p, y)] - \mathbb{E}[\ell_v^+(p + \varepsilon, y)] = \frac{\varepsilon}{2} - \left(-\frac{\varepsilon}{2}\right) = \varepsilon.$$

Therefore,  $\text{CDL}_{\mathcal{M}_+}(J_1) \gtrsim \text{CE}_{\text{Int}}(J_1)$ .

For  $J_2$ , consider a uniform  $p \sim [0, 1 - \varepsilon]$ , and suppose that  $y|p \sim \text{Ber}(p + \varepsilon)$ . Then,

$$\text{CE}_{\text{Int}}(J_2) \geq \mathbb{E}[y - p] = \varepsilon,$$

but the characterization of CDL in [Corollary 3.6](#) implies

$$\begin{aligned} \text{CDL}_{\mathcal{M}_+}(J_2) &= \sup_{\substack{v \in [0, 1], \\ \kappa \in \mathcal{M}_+}} \mathbb{E}[\ell_v^+(p, y)] - \mathbb{E}[\ell_v^+(p, y)] \\ &= \sup_{v \in [0, 1]} \mathbb{E}[\ell_v^+(p, y)] - \mathbb{E}[\ell_v^+(p + \varepsilon, y)] \\ &= \sup_{v \in [0, 1]} \mathbb{E}[(\text{sign}_+(p + \varepsilon - v) - \text{sign}_+(p - v))(p + \varepsilon - v)] \\ &= \sup_{v \in [0, 1]} 2 \cdot \mathbb{E}[\mathbf{1}[v - \varepsilon \leq p < v](p + \varepsilon - v)] \\ &\leq \sup_{v \in [0, 1]} 2 \cdot \mathbb{E}[\mathbf{1}[v - \varepsilon \leq p < v](v + \varepsilon - v)] \\ &= 2 \cdot \frac{\varepsilon}{1 - \varepsilon} \cdot \varepsilon \\ &\lesssim \varepsilon^2. \end{aligned}$$

Thus,  $\text{CDL}_{\mathcal{M}_+}(J_2) \lesssim \text{CE}_{\text{Int}}(J_2)^2$ .  $\square$



## D Properties of Generalized Monotone Post-Processings

We provide here some classical results related to generalized monotone post-processings, which, in particular, imply [Corollary 6.2](#). We begin with the following result on their VC dimension.

**Proposition D.1.** *For any  $r \in \mathbb{N}$ ,  $\text{VCdim}(\text{thr}(\mathcal{M}_r)) = 2r$ .*

*Proof.* Consider the points  $p_i = \frac{i}{2r}$  where  $i \in [2r]$ . We will show the following:

$$\{\mathbf{v} \in \{\pm 1\}^{2r} : \mathbf{v} = (\text{sign}_+(\kappa(p_i) - 1/2))_{i \in [2r]} \text{ for some } \kappa \in \mathcal{M}_r\} = \{\pm 1\}^{2r}$$

In particular, for any  $\mathbf{v} \in \{\pm 1\}^{2r}$ , we form the intervals  $I_1, I_2, \dots, I_r$  as follows. Let  $I_1$  be an interval of the form  $[p_i, p_j]$ , where  $i \leq j$  and  $\mathbf{v}(i) = \mathbf{v}(i+1) = \dots = \mathbf{v}(j) = 1$  and  $\mathbf{v}(i') = -1$  for all  $i' < i$  and  $i' = j+1$ . We then proceed recursively to form  $I_2, \dots, I_r$  after removing the points  $p_1, \dots, p_j, p_{j+1}$ . Note that for each interval we form, we remove at least 2 points. One for  $p_j$  and one for  $p_{j+1}$ . We let  $\kappa_{\mathbf{v}}(p) = \mathbf{1}[p \in \cup_{i \in [r]} I_i]$ , where we have  $\kappa_{\mathbf{v}} \in \mathcal{M}_r$  and  $(\text{sign}_+(\kappa(p_i) - 1/2))_{i \in [2r]} = \mathbf{v}$ .

On the other hand, for any set of  $2r+1$  distinct points on  $[0, 1]$ , no function in  $\mathcal{M}_r$  can generate the labeling  $\mathbf{v}' \in \{\pm 1\}^{2r+1}$ , where  $\mathbf{v}'(2i-1) = 1$  and  $\mathbf{v}'(2i) = -1$ , for  $i = 1, 2, \dots, r$ , and  $\mathbf{v}'(2r+1) = 1$ , because the set  $\{p : \kappa(p) \geq 1/2\}$  would then have at least  $r+1$  disjoint components.  $\square$

Our testing result in [Corollary 6.2](#), which pertains to the specific class  $\mathcal{M}_r$ , is achieved by instantiating our more general [Theorem 6.1](#) with the following standard agnostic learner for unions of  $r$  intervals, based on dynamic programming (see Section 4.2 in [\[KSS94\]](#)). Indeed, the tester of [Theorem 6.1](#) makes  $O(\log(1/\varepsilon\delta)/\varepsilon)$  non-adaptive calls to the agnostic learner, which has sample complexity  $O((r + \log 1/\delta)/\varepsilon^2)$  and runtime  $\tilde{O}(r^2 + r \log 1/\delta)/\varepsilon^2$ . Therefore, a union bound over the failure probability of each call implies a tester with total sample complexity  $\tilde{O}(r/\varepsilon^2)$  and runtime  $\tilde{O}(r^2/\varepsilon^3)$ . For completeness, we describe and analyze the proper agnostic learner here.

**Theorem D.2** ([\[KSS94\]](#)). *Let  $r \geq 1$  and  $\mathcal{C} = \text{thr}(\mathcal{M}_r)$ . For any  $\varepsilon, \delta \in (0, 1)$ , there is a proper agnostic  $(\varepsilon, \delta)$ -learner for  $\mathcal{C}$  with sample complexity  $O((r + \log 1/\delta)/\varepsilon^2)$  and runtime  $\tilde{O}((r^2 + r \log 1/\delta)/\varepsilon^2)$ .*

*Proof of Theorem D.2.* We will show that there is an algorithm  $\text{ERM}(\mathcal{C})$  that takes as input a set  $S$  of labeled examples of the form  $(p, z)$  where  $p \in [0, 1]$  and  $z \in \{\pm 1\}$ , runs in time  $O(|S|r + |S| \log |S|)$ , and outputs some  $h \in \mathcal{C}$  such that the following holds:

$$\mathbb{P}_{(p,z) \sim S}[h(p) \neq z] \leq \min_{f \in \mathcal{C}} \mathbb{P}_{(p,z) \sim S}[f(p) \neq z]$$

In [Algorithm 3](#) we present a version of  $\text{ERM}(\mathcal{C})$  that does not return  $h$ , but only returns an estimate of its error. The algorithm  $\text{ERM}(\mathcal{C})$  can be implemented based on [Algorithm 3](#) by using some additional space. The agnostic learning result then follows by standard uniform convergence arguments, due to the fact that  $\text{VCdim}(\mathcal{C}) = 2r$  ([Proposition D.1](#)), as long as  $m := |S| \geq C(r + \log(1/\delta))/\varepsilon^2$  for some sufficiently large constant  $C \geq 1$ .

We first observe that  $\mathbf{1}[z \neq f(p)] = (1 - z \cdot f(p))/2$  for any  $f \in \mathcal{C}, z \in \{\pm 1\}$ . Moreover, any  $f \in \mathcal{C}$  can be expressed as follows for  $r$  disjoint intervals  $I_1, I_2, \dots, I_r \subseteq [0, 1]$ :

$$f(p) = 2 \sum_{i=1}^r \mathbf{1}[p \in I_i] - 1$$

---

**Algorithm 3:** ESTIMATEMINIMUMRISK( $S, r$ )

---

**Input:** Set  $S$  of  $m$  pairs of the form  $(p, z)$  where  $p \in [0, 1]$ ,  $z \in \{\pm 1\}$  and  $r \geq 1$ .

**Output:** A value  $R \in [0, 1]$ .

```
/* Sorting the Samples, in time:  $O(m \log(m))$  */
1 Let  $\mathcal{O} = ((p_1, z_1), (p_2, z_2), \dots, (p_m, z_m))$  be an ordering such that  $p_1 \leq p_2 \leq \dots \leq p_m$ ;
2 If there are any duplicates  $p_i = p_{i+1} = \dots = p_{i+k}$ , then merge them and set  $z' = \sum_{j=0}^k z_{i+j}$ ,
   so that we obtain  $\mathcal{O}' = ((p'_1, z'_1), (p'_2, z'_2), \dots, (p'_{m'}, z'_{m'}))$  with  $p'_1 < p'_2 < \dots < p'_{m'}$ ;
/* Initializations, in time:  $O(m)$  */
3 Set  $Z(i) = \sum_{j \leq i} z'_j$ ,  $Z(0) = 0$ ;
4 Set  $Q(0, t) = Q(s, 1) = 0$  for all  $t \in [m' + 1]$  and  $s \in [r]$ ;
5 Set  $M(0, j) = \max_{1 \leq i \leq j} \{-Z(i - 1)\}$  for all  $j \in [m' + 1]$  and  $M(s, 1) = 0$  for all  $s \in [r]$ ;
6 Set  $B(0, t) = \max_{1 \leq j < t} \{Z(j) + M(0, j)\}$  for all  $t \in [m' + 1]$  and  $B(s, 1) = 0$  for all  $s \in [r]$ ;
/* Dynamic Programming, in time:  $O(mr)$  */
7 for  $s = 1, 2, \dots, r$  do
8   for  $t = 2, 3, \dots, m' + 1$  do
9      $Q(s, t) = \max\{Q(s - 1, t), B(s - 1, t)\}$ ;
10  for  $j = 2, 3, \dots, m' + 1$  do
11     $M(s, j) = \max\{M(s, j - 1), Q(s - 1, j) - Z(j - 1)\}$ ;
12     $B(s, j) = \max\{B(s, j - 1), Z(j) - M(s, j)\}$ ;
13 Let  $R = Q(r, m' + 1)/m$ ;
```

---

We therefore have the following:

$$\begin{aligned} \min_{f \in \mathcal{C}} \mathbb{P}_{(p,z) \sim S}[f(p) \neq z] &= \frac{1}{2} - \frac{1}{2} \max_{f \in \mathcal{C}} \mathbb{E}_{(p,z) \sim S}[zf(p)] \\ &= \frac{1}{2} + \frac{1}{2} \mathbb{E}_{(p,z) \sim S}[z] - \frac{1}{2} \max_{I_1, \dots, I_r} \sum_{i=1}^r \mathbb{E}_{(p,z) \sim S}[z \mathbf{1}[p \in I_i]] \end{aligned}$$

It suffices to find the endpoints of the intervals  $(I_i)_i$  that maximize the following quantity:

$$Q := \max_{I_1, \dots, I_r} \sum_{i=1}^r \sum_{(p,z) \in S} z \mathbf{1}[p \in I_i]$$

Let  $S'$  be the set of pairs of the form  $(p, z')$ , where each  $p \in [0, 1]$  appears in  $S$  at least once and  $z'$  is the sum of all  $z$  such that  $(p, z) \in S$ . Each  $p \in [0, 1]$  appears in  $S'$  at most once. Let  $P$  be the set of  $p \in [0, 1]$  appearing in  $S'$ . Then, we can write  $Q$  as follows:

$$Q = \max_{k \in [r]} \max_{\substack{a_i, b_i \in P \\ a_i \leq b_i < a_{i+1}}} \sum_{i=1}^k \sum_{(p,z') \in S'} z' \mathbf{1}[p \in [a_i, b_i]].$$

Let  $P(q) = P \cap [0, q]$  and let  $Q(k, q)$  be defined as follows:

$$Q(k, q) := \max_{k' \leq k} \max_{\substack{a_i, b_i \in P(q) \\ a_i \leq b_i < a_{i+1}}} \sum_{i=1}^{k'} \sum_{(p,z') \in S'} z' \mathbf{1}[p \in [a_i, b_i]].$$

Then,  $Q(k, q)$  satisfies the following recurrence for all  $k \in [r]$  and  $q \in P \cup \{\infty\}$ :

$$Q(k, q) = \max \left\{ Q(k-1, q), \max_{\substack{a, b \in P(q) \\ a \leq b}} \left\{ Q(k-1, a) + \sum_{(p, z') \in S'} z' \mathbf{1}[p \in [a, b]] \right\} \right\},$$

as long as the initial conditions are the following for all  $q \in P \cup \{\infty\}$  and  $k \in [r]$ :

$$Q(0, q) = Q(k, p_{\min}) = 0, \text{ where } p_{\min} = \min_{p \in P} p$$

Our goal is to estimate the quantity  $Q = Q(r, \infty)$ , and retrieve the points  $(a_i, b_i)_i$  that achieve the maximum. Due to the structure of  $Q(k, q)$ , the estimation of  $Q$  can be achieved in time  $\text{poly}(r, |P|)$  via dynamic programming. For the indices  $(a_i, b_i)_i$ , we associate each pair  $(k, q)$  with a set of intervals  $\mathcal{I}(k, q)$  of the form  $[a, b]$ , so that:

$$\mathcal{I}(k, q) = \begin{cases} \mathcal{I}(k-1, q), & \text{if } Q(k, q) = Q(k-1, q) \\ \mathcal{I}(k-1, a) \cup \{[a, b]\}, & \text{if } Q(k, q) = Q(k-1, a) + \sum_{(p, z') \in S'} z' \mathbf{1}[p \in [a, b]] \end{cases}$$

In order to obtain an improved time complexity, we may use some additional memory to store intermediate auxiliary quantities as described in [Algorithm 3](#). In particular, we define the following quantities for all  $a \in P$ :

$$Z(a) = \sum_{(p, z') \in S'} z' \mathbf{1}[p \leq a], \text{ and } \text{prev}(a) = \max_{a' \in P(a)} a'$$

and we equivalently formulate  $Q(k, q)$  as follows:

$$\begin{aligned} Q(k, q) &= \max \left\{ Q(k-1, q), \max_{\substack{a, b \in P(q) \\ a \leq b}} \left\{ Q(k-1, a) + Z(b) - Z(\text{prev}(a)) \right\} \right\} \\ &= \max \left\{ Q(k-1, q), \max_{b \in P(q)} \left\{ Z(b) + \max_{a \leq b} \left\{ Q(k-1, a) - Z(\text{prev}(a)) \right\} \right\} \right\} \end{aligned}$$

where we may define the following quantities:

$$\begin{aligned} M(k-1, b) &= \max_{a \leq b} \left\{ Q(k-1, a) - Z(\text{prev}(a)) \right\} \\ B(k-1, q) &= \max_{b \in P(q)} \left\{ Z(b) + M(k-1, b) \right\} \end{aligned}$$

Due to the definitions of the auxiliary quantities, they satisfy the following recurrence relations.

$$\begin{aligned} Q(k, q) &= \max\{Q(k-1, q), B(k-1, q)\} \\ B(k, q) &= \max\{B(k, \text{prev}(q)), Z(q) + M(k, q)\} \\ M(k, q) &= \max\{M(k, \text{prev}(q)), Q(k-1, q) - Z(\text{prev}(q))\} \end{aligned}$$

Each step of the above recurrence relations can be computed in  $O(1)$ , which implies the desired bound on the runtime.  $\square$