




# Smooth Calibration and Decision Making

Jason Hartline   

Department of Computer Science, Northwestern University, Evanston, IL, USA

Yifan Wu   

Department of Computer Science, Northwestern University, Evanston, IL, USA

Yunran Yang  

Zhiyuan College, Shanghai Jiao Tong University, China

---

## Abstract

Calibration requires predictor outputs to be consistent with their Bayesian posteriors. For machine learning predictors that do not distinguish between small perturbations, calibration errors are continuous in predictions, e.g. smooth calibration error [9], distance to calibration [2]. On the contrary, decision-makers who use predictions make optimal decisions discontinuously in probabilistic space, experiencing loss from miscalibration discontinuously. Calibration errors for decision-making are thus discontinuous, e.g., Expected Calibration Error [10], and Calibration Decision Loss [19]. Thus, predictors with a low calibration error for machine learning may suffer a high calibration error for decision-making, i.e. they may not be trustworthy for decision-makers optimizing assuming their predictions are correct. It is natural to ask if post-processing a predictor with a low calibration error for machine learning is without loss to achieve a low calibration error for decision-making. In our paper, we show post-processing an online predictor with  $\epsilon$  distance to calibration achieves  $O(\sqrt{\epsilon})$  ECE and CDL, which is asymptotically optimal. The post-processing algorithm adds noise to make predictions differentially private. The optimal bound from low distance to calibration predictors from post-processing is non-optimal compared with existing online calibration algorithms that directly optimize for ECE and CDL.

**2012 ACM Subject Classification** Theory of computation  $\rightarrow$  Machine learning theory; Theory of computation  $\rightarrow$  Algorithmic game theory and mechanism design; Theory of computation  $\rightarrow$  Online learning algorithms

**Keywords and phrases** Calibration, calibration errors, decision making, differential privacy

**Digital Object Identifier** 10.4230/LIPIcs.FORC.2025.16

**Related Version** *Full Version:* <https://arxiv.org/abs/2504.15582>

## 1 Introduction

Calibration requires that predictions are empirically conditionally unbiased. Consider a sequence of predictions for the chance of rain, a predictor is calibrated if, for every  $p$ , among the days that the prediction is  $p$ , the fraction of rainy days is also  $p$ . Calibrated predictions can thus be reliably interpreted as probabilities.

Calibration errors quantify the error of a predictor from being perfectly calibrated. Machine learning (ML) predictors make predictions continuously in probabilistic space, so calibration errors for ML are continuous in prediction values and do not distinguish between small perturbations in predictions. Two canonical examples are the *smooth calibration error* [9] and the *distance to calibration* (DTC) [2]. As an illustrating example of the calibration errors for ML, consider a predictor in Table 1. Although the predictions of 50.01% and 49.99% are biased, the total number of rainy days is 50%, indicating the predictor is very close to a calibrated predictor that always outputs 50%. Both DTC and the smooth calibration error are about 0.01%, close to 0. The smooth calibration error combines the bias over all the days by weighing biases continuously, e.g. weighing bias  $(50.01\% - 0)$  by  $-0.01\%$ ,  $(49.99\% - 1)$  by  $0.01\%$ , and summing together (the weights are Lipschitz continuous in prediction values).



© Jason Hartline, Yifan Wu, and Yunran Yang;

licensed under Creative Commons License CC-BY 4.0

6th Symposium on Foundations of Responsible Computing (FORC 2025).

Editor: Mark Bun; Article No. 16; pp. 16:1–16:26

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 16:2 Smooth Calibration and Decision Making

■ **Table 1** A miscalibrated predictor for the chance of rain.

| Prediction value | # days           | conditional frequency of rain |
|------------------|------------------|-------------------------------|
| 50.01%           | half of the days | 0                             |
| 49.99%           | half of the days | 1                             |

The smooth calibration error is linearly related to DTC, which calculates the expected  $\ell_1$  distance between the predictor and the nearest calibrated predictor, which in this example predicts 50% every day.

Decision-makers make decisions discontinuously in probabilistic space, thus, a calibration error for decision-making is discontinuous in the prediction space. For example, consider a decision problem with binary action space, bringing an umbrella or not. The decision maker receives a payoff of 1 when the decision matches the state, i.e. bringing an umbrella when rainy, not bringing when not rainy, and a payoff of 0 in other cases. When assisted by a prediction, the action of a decision-maker changes from not bringing an umbrella to bringing an umbrella at the prediction threshold of 50%. Two examples of calibration errors for decision-making are Expected Calibration Error (ECE) [10] and Calibration Decision Loss (CDL) [19]. CDL quantifies the worst-case decision loss of a decision-maker who trusts the prediction as a probability, where the worst-case is taken over all payoff-bounded decision tasks. By definition, CDL upperbounds any decision-maker's loss. ECE, the most well-studied calibration error metric, is defined by the averaged absolute bias in predictions. For example, ECE averages over  $|50.01\% - 0|$  and  $|49.99\% - 1|$  for the predictor in Table 1 and has a calibration error of 50.01%. Kleinberg et al. [23] shows that ECE linearly upperbounds the decision loss of every payoff-bounded decision task, implying an upperbound of CDL.

From the decision-making perspective, having a low calibration error for ML, however, does not guarantee a low calibration error for decision-making or being trustworthy for decision-making. Consider the same example of a predictor in Table 1 and the umbrella decision problem above. According to a calibration error for ML, e.g. distance to calibration, the predictor is 0.01% close to a calibrated predictor that always outputs 50%. However, to the decision-maker, the prediction suggests not taking an umbrella when the weather is rainy, and taking an umbrella when not rainy. This non-trustworthiness comes from the discontinuity of decision-making which the decision-maker changes an action at the threshold 50%.

Here is the natural question: can we design a post-processing algorithm that, given any predictor with a low calibration error for machine learning, outputs predictions with a low calibration error for decision-making? Ideally, the post-processing algorithm should achieve near-optimal guarantees that asymptotically match the guarantees from directly optimizing for decision-making.

Our paper designs a post-processing algorithm that, given any predictor with  $DTC = \epsilon$ , outputs differentially private predictions with ECE and CDL bounded by  $O(\sqrt{\epsilon})$ , in both the batch setting and the online setting. We give lower bounds, described below, for both that online and batch setting, that show that this post-processing algorithm is asymptotically optimal. Additionally the online lower bounds shows that the optimal predictors for decision makers cannot be constructed from optimal predictors from machine learning.

We show that the privacy-based post-processing algorithm is asymptotically optimal in the online setting. This optimality implies there does not exist a post-processing algorithm that achieves the same guarantee as known online algorithms that directly optimize predictions for ECE and CDL. For online calibration, there has been shown an  $O\left(T^{-\frac{1}{3}-c}\right)$  ( $c > 0$ )

upperbound on optimal algorithm for ECE [5], a  $\tilde{O}(T^{-\frac{1}{2}})$  optimal bound to CDL [19], and an  $\Omega(T^{-\frac{2}{3}})$  lowerbound to DTC [27]. Thus, applying the lowerbound of  $\Omega(\sqrt{\epsilon})$ , any post-processing algorithm can only achieve the non-optimal  $\Omega(T^{-\frac{1}{3}})$  ECE and CDL.

We show that the privacy-based post-processing algorithm is asymptotically optimal in the batched setting in two models. The first model considers post-processing algorithms applied individually to each prediction, and the same guarantee and lowerbound to ECE and CDL applies as in the online setting. The second model allows algorithms that post-process the entire batch of predictions. However, doing so just to attain calibration is too easy: simply ignoring the individual information in each prediction and averaging them all will be close to calibrated. Thus, we impose a stronger benchmark that measures the worst-case decision loss relative to a nearby – in the sense of  $\epsilon$  Distance to Calibration – calibrated predictor. This worst case is taken over all such nearby calibrated predictors and all bounded decision problems. We show that the privacy-based post-processing algorithm achieves  $O(\sqrt{\epsilon})$  decision loss and that this result is tight, i.e. no other post-processing algorithm achieves asymptotically better decision loss.

## 1.1 Related Work

### Calibration Error Metrics

The most relevant work to ours, Błasiok and Nakkiran [4], introduces the error metric Smooth ECE, which, given a predictor, calculates the ECE with Gaussian noise added to the predictions. For any predictor with  $\text{DTC} = \epsilon$ , smooth ECE is shown to be bounded by  $\Theta(\sqrt{\epsilon})$ . Instead, our paper focuses on the decision-making perspective of calibration. We show that this bound of  $\Theta(\sqrt{\epsilon})$  is tight, suggesting that from a decision-making perspective, optimizing for DTC and post-processing achieves suboptimal guarantees. Our post-processing algorithm also generalizes the result of Błasiok and Nakkiran [4] by considering noise distributions for differential privacy.

As introduced previously, existing calibration error metrics mainly focus on two aspects: calibration errors for machine learning, continuous in predictions, e.g. smooth calibration error [9], distance to calibration<sup>1</sup> [2], smooth ECE [20]; and calibration errors for decision-making, e.g. the canonical ECE [10] and the Calibration Decision Loss [19]. Recently, as an orthogonal property to continuity and decision-making, Haghtalab et al. [16] propose an approximately truthful calibration error metric for an expected-error-minimizing sequential predictor.

### Online Calibration

In online calibration, the predictor repeatedly interacts with an adversary selecting a binary state. In each round, both the predictor and the adversary know the history of predictions and states, but are not allowed to strategize conditioned on the opponent’s action in the current round. Foster and Vohra [11] showed an upperbound of  $\text{ECE} = O(T^{-\frac{1}{3}})$ , which is recently proven to be polynomial-time achievable by Noarov et al. [25]. Recently, Dagan et al. [7] improves the upperbound to  $O\left(T^{-\frac{1}{3}-c}\right)$  for some constant  $c > 0$ . On the lowerbound side, Qiao and Valiant [26] showed there exists an  $O(T^{-0.472})$  lowerbound, strictly above  $\tilde{O}\left(\frac{1}{\sqrt{T}}\right)$ , which is improved to  $O(T^{-0.456})$  by Dagan et al. [6].

<sup>1</sup> We follow Qiao and Zheng [27] and refer to *distance to calibration* as the *lower* distance to calibration in Błasiok et al. [2].

## 16:4 Smooth Calibration and Decision Making

For linearly related smooth calibration error and DTC, Qiao and Zheng [27] prove an  $O(\frac{1}{\sqrt{T}})$  upperbound and an  $O(T^{-\frac{2}{3}})$  lowerbound. Arunachaleswaran et al. [1] design a simple polynomial-time algorithm that achieves  $\text{DTC} = O(\frac{1}{\sqrt{T}})$ .

The Calibration Decision Loss (CDL) is introduced in Hu and Wu [19] with a bound of  $\tilde{O}(\frac{1}{\sqrt{T}})$ , tight up to a logarithmic factor.

### Omniprediction

Our definition of decision loss for the batch setting can be equivalently formulated as achieving omniprediction with regard to reference predictors and a set of loss functions. Calibration guarantees the trustworthiness of predictions by every decision-maker, allowing decision-making to be separated from predictions. Introduced in Gopalan et al. [13], omnipredictor follows the same idea, requiring an omnipredictor to achieve comparable guarantee with regard to a class of loss functions and a set of competing predictors. Techniques from the algorithmic fairness literature, e.g. Hebert-Johnson et al.; Kim et al. [17, 22], have been applied to achieve omniprediction in both online and batch setting [13, 15, 14, 12, 18]. While the classical guarantee usually learns an omnipredictor that competes with the hypothesis space of predictors, our decision loss evaluates a predictor with regard to the set of calibrated predictors close in DTC.

## 2 Preliminaries

### Mathematical Notations

We write  $D_{X,Y}$  as the joint distribution between random variables  $X$  and  $Y$ , and  $X \sim D$  as random variable  $X$  drawn from distribution  $D$ . Where it is obvious from the context, we write  $\Pr[X = x]$  for the probability of a discrete random variable as well as the probabilistic density function of a continuous random variable.

We consider a prediction problem of a binary state  $\theta \in \Theta = \{0, 1\}$ . A predictor is specified by a joint distribution  $D_{P,\Theta}$  over the prediction  $p$  and the state  $\theta$ . Slightly abusing the notation, we also write a predictor as a random variable  $P$ , omitting the state, where a realized prediction value is  $p$ .

Our privacy-based post-processing algorithm adds noise to make predictions differentially private. Definition 1 defines a differentially private mechanism for predictions.

► **Definition 1** (Differential Privacy). *A mechanism  $\mathcal{M}$  is  $(\gamma, \delta)$ -differentially private (DP) if for any two predictions  $q, q' \in [0, 1]$ :*

$$\Pr[\mathcal{M}(q) \in \mathcal{I}] \leq e^{\gamma|q-q'|} \cdot \Pr[\mathcal{M}(q') \in \mathcal{I}] + \delta.$$

We construct our privacy-based algorithm by adding truncated noise, where truncation guarantees predictions fall in the range of  $[0, 1]$ . The truncation of noise  $Y$  works in the following way: given a prediction  $q$ , for random variable  $Y$  with unbounded support, we draw  $X \sim D_\epsilon(q)$  such that

$$\Pr[q + X = p] = \frac{\Pr[q + Y = p]}{\Pr[q + Y \in [0, 1]]}.$$

## 2.1 Predictions for Decision-Making

A decision maker faces a decision problem  $(A, \Theta, U)$ :

- the decision maker (DM) selects an action  $a \in A$ ;
- a payoff-relevant state  $\theta \in \Theta$  is realized;
- DM receives payoff  $U : A \times \Theta \rightarrow \mathbb{R}$ .

When assisted with a prediction, the best response  $a^*$  maps a prediction to the action:

$$a^*(p) = \arg \max_{a \in A} \mathbf{E}_{\theta \sim p} [U(a, \theta)]. \quad (1)$$

When the DM best responds, she assumes the state is drawn from  $p$  and takes the action that maximizes the expected payoff. We define the best-responding payoff as a function  $S$  of the prediction and the state.

► **Definition 2** (Scoring Rule from Decision). *Given a decision problem  $U$ , the scoring rule induced from  $U$  and belief  $p \in \Delta(\Theta)$  is*

$$S_U(p, \theta) = U(a^*(p), \theta)$$

Proper scoring rules characterize scoring rules induced from a decision problem.

► **Definition 3** (Proper Score). *A scoring rule  $S : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}$  is proper if and only if*

$$\mathbf{E}_{\theta \sim p} [S(p, \theta)] \geq \mathbf{E}_{\theta \sim p} [S(p', \theta)], \forall p' \in [0, 1].$$

Claim 4 shows that the space of best-responding payoff is equivalent to the space of proper scoring rules. Throughout the paper, we will write the best-responding decision payoff as proper scoring rules.

▷ **Claim 4** (Kleinberg et al. [23]; Hu and Wu [19]). *There exists a bijective mapping between a bounded proper scoring rules and scoring rule induced from a decision problem with bounded payoff:*

- Given a decision problem  $U(\cdot, \cdot) \in [0, 1]$ , the induced scoring rule  $S_U(\cdot, \cdot) \in [0, 1]$  is a proper scoring rule.
- Given a proper scoring rule  $S(\cdot, \cdot) \in [0, 1]$ , there exists a decision problem  $U(\cdot, \cdot) \in [0, 1]$  that induces  $S$ .

Given a set of reference predictors  $\mathcal{B}$  and a set of proper scoring rules  $\mathcal{S}$ , we define the decision loss with regard to the set of reference predictors.

► **Definition 5** (Decision Loss). *Given a set of reference predictors  $\mathcal{B}$  and a set of proper scoring rules  $\mathcal{S}$ , the decision loss of a predictor  $P$  is*

$$DL(P; \mathcal{B}) = \max_{S \in \mathcal{S}, B \in \mathcal{B}} \mathbf{E}_{p, b, \theta \sim D_{P, B, \Theta}} [S(b, \theta) - S(p, \theta)].$$

Throughout the paper, we consider decision loss with regard to the set of all bounded proper scoring rules  $\mathcal{S} = \{S(\cdot, \cdot) \in [0, 1]\}$ , i.e. all decision problems with bounded payoff.

Our decision loss is closely related to omniprediction. A predictor with  $\epsilon$  decision loss is an  $\epsilon$  omnipredictor with regard to reference predictors in  $\mathcal{B}$  and the set of scoring rules in  $\mathcal{S}$ .

► **Definition 6** (Omniprediction [13]). *Given a set of reference predictors  $\mathcal{B}$  and a set of proper scoring rules  $\mathcal{S}$ , a predictor is an  $\epsilon$ -omnipredictor with regard to  $\mathcal{B}$  and  $\mathcal{S}$  if*

$$\mathbf{E}_{(p, \theta) \sim D_{P, \Theta}} [S(p, \theta)] \geq \mathbf{E}_{(b, \theta) \sim D_{B, \Theta}} [S(b, \theta)] - \epsilon, \quad \forall B \in \mathcal{B}, S \in \mathcal{S}.$$

## 2.2 Measures of Calibration Error

In this section, we define different calibration error metrics that are relevant to the paper. The definitions of error metrics follow the definitions of perfect calibration. We denote the Bayesian posterior of prediction values as  $\hat{p} = \Pr[\theta = 1|P = p]$ .

► **Definition 7** (Perfect Calibration). *A predictor  $P$  is perfectly calibrated if  $p = \hat{p}$  for any  $p \in [0, 1]$ .*

We introduce relevant calibration errors to the paper by two categories: calibration error for decision-making and calibration error for machine learning.

### 2.2.1 Calibration Errors for Decision-Making

The canonical calibration error metric is ECE, the averaged bias in predictions.

► **Definition 8** (Expected Calibration Error, ECE). *Given predictor  $P$ , the expected calibration error is*

$$ECE(P) = \mathbf{E}_{p \sim P} [|p - \hat{p}|].$$

The swap regret of a decision-maker is closely related to predictions being calibrated. Swap regret minimizers are special cases of omnipredictors where the set of reference predictors  $\mathcal{B}$  is the set of post-processed predictors, i.e. by applying a mapping  $\sigma : [0, 1] \rightarrow [0, 1]$  to the original predictions  $p$ .

► **Definition 9** (Swap Regret). *Given a decision problem with proper scoring rule  $S$ , a predictor  $P$ , the swap regret for the decision-maker is*

$$SWAP_S(P) = \max_{\sigma: [0,1] \rightarrow [0,1]} \mathbf{E}_{(p,\theta) \sim D_{p,\theta}} [S(\sigma(p), \theta) - S(p, \theta)].$$

Proposition 10 shows that the swap regret equals the payoff improvement from calibrating a predictor.

► **Proposition 10.** *Given a decision problem with proper scoring rule  $S$ , a predictor  $P$ , the mapping  $\sigma^*(p) = \Pr[\theta|p]$  is the swap regret maximizing mapping, i.e.*

$$\sigma^* \in \arg \max_{\sigma: [0,1] \rightarrow [0,1]} \mathbf{E}_{(p,\theta) \sim D_{p,\theta}} [S(\sigma(p), \theta) - S(p, \theta)].$$

*the swap regret equals the payoff improvement from calibrating the predictor:  $\sigma^*(p) = \Pr[\theta = 1|p]$ .*

Proposition 11 thus follows by the definition of calibration.

► **Proposition 11** (Foster and Vohra [11]). *A predictor is calibrated if and only if for any decision problem  $U$ , the decision-maker has no swap regret.*

Motivated by Proposition 11, Hu and Wu [19] define Calibration Decision Loss (CDL), the worst-case decision loss induced by miscalibration, with worst-case over all bounded proper scoring rules. Instead of decision loss that compares with a set of fixed reference predictor, CDL calculates the decision loss where the reference is the calibrated correspondence of the predictor to be evaluated.

► **Definition 12** (Calibration Decision Loss, CDL). For predictor  $P$ , the Calibration Decision Loss is defined as

$$CDL(P) = \max_{\text{proper } S(\cdot, \cdot) \in [0,1]} SWAP_S(P),$$

where the maximum is taken over all bounded proper scoring rules.

Kleinberg et al. [23] shows that CDL is upperbounded by ECE.

► **Lemma 13** (Kleinberg et al. [23]). For predictor  $P$ , CDL is upperbounded by ECE, i.e.  $CDL(P) \leq ECE(P)$ .

## 2.2.2 Calibration Errors for Machine Learning

The calibration errors in this section are continuous in prediction space. The calibration error metrics relevant to the paper are the smooth calibration error and the distance to calibration (DTC).

► **Definition 14** (Smooth Calibration Error). Given predictor  $P$ , the smooth calibration error takes the supremum over the set  $\Sigma$  of 1-Lipschitz functions:

$$SMCAL(P) = \sup_{\sigma \in \Sigma} \mathbf{E}_{(p, \theta) \sim D_{P, \Theta}} [\sigma(p) \cdot (p - \theta)]$$

Note that without the constraint that  $\sigma$  is 1-Lipschitz, SMCAL is the same as ECE. To see this, note that taking  $\sigma = 1$  when  $p - \hat{p} \geq 0$  and  $\sigma = -1$  when  $p - \hat{p} < 0$  gives the same definition as ECE.

Given a predictor  $p$ , the distance to calibration finds a calibrated predictor  $b$  with a coupling  $D_{p, b, \theta}$  with the given predictor  $p$ , such that  $b$  has the smallest distance from  $p$ . The distance is the  $\ell_1$  distance between predictions, as defined in Definition 15.

► **Definition 15** (Distance between Predictors). Given predictors  $B$  and  $P$ , the distance between the predictors is defined as

$$DIST(P, B) = \mathbf{E}_{D_{P, B}} [|P - B|].$$

Definition 16 defines the distance to calibration.

► **Definition 16** (Distance to Calibration, DTC). For predictor  $P$ , the distance to calibration is

$$DTC(P) = \min_{B \text{ is calibrated}} DIST(P, B).$$

The smooth calibration error is linearly related to the distance to calibration. While in our paper, we mainly focus on DTC, our results also apply to the smooth calibration error SMCAL.

► **Lemma 17** (Błasiok et al. [2]). Given any predictor  $P$ ,

$$\frac{1}{2} DTC(P) \leq SMCAL(P) \leq DTC(P).$$

## 2.3 Online and Batch Post-Processing Algorithm

We design a post-processing algorithm for both the online setting and the batch setting, given predictions  $q_1 \dots q_T$  from a predictor  $Q$ . The post-processing algorithm knows the parameter  $DTC(Q) = \epsilon$ .

### The Online Setting

In the online setting, the goal of a post-processing algorithm is to generate trustworthy predictions  $\mathbf{p} = (p_1, \dots, p_T)$  with low ECE or CDL given a sequence of predictions with low DTC. At the end of  $T$  rounds, the predictor is evaluated by a calibration error against the sequence of states  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_T)$ . We define the joint distribution of  $D_{P,\Theta}$  in definitions in Section 2.2 as the empirical distribution of  $(p_t, \theta_t)$  over  $T$  rounds, which gives equivalent definitions of online calibration errors in the literature. We will write the calibration error of online predictors as a function of  $\mathbf{p}$  and  $\boldsymbol{\theta}$ .

In round  $t \in [T]$ , the adversary selects a prediction  $q_t$ . The post-processing algorithm  $f = (f_t)_{t \in [T]}$  makes a (randomized) prediction according to  $f_t$  given  $q_t$  and the history of  $(q_k, p_k)_{k \in [t-1]}$  but not the states<sup>2</sup>. The adversary then reveals the state  $\theta_t$ . The adversary knows the full history of interactions, i.e.  $(q_k, p_k, \theta_k)_{k \in [t-1]}$ . When selecting the prediction  $q_t$ , the adversary faces the constraint that  $\text{DTC}(Q) = \epsilon$  at the end of  $T$  rounds.

Note that the restriction of the algorithm not knowing the state is slightly different from the classic online calibration [11]. This restriction effectively excludes a post-processing algorithm that ignores the predictions  $q$  and directly implements a calibrated predictor.

### The Batch Setting

In the batch setting, the predictor  $Q$  is specified by the joint distribution  $D_{Q,\Theta}$  as introduced in the beginning of Section 2. We write  $Q^T$  as the joint distribution of  $T$  independent and identical draws of predictions from  $Q$ . Given  $T$  realizations of predictions  $\mathbf{q} = (q_1, \dots, q_T) \sim Q^T$ , the post-processing algorithm  $f : [0, 1]^T \rightarrow \Delta([0, 1]^T)$  outputs (randomized) predictions  $\mathbf{p} = (p_1, \dots, p_T)$ . Since  $f$  is only allowed to depend on predictions  $\mathbf{q}$  not the states, it is without loss to write  $f_{\mathbf{q}}(q) : [0, 1] \rightarrow \Delta([0, 1])$ , assuming the output follows the same distribution fixing samples  $\mathbf{q}$ . Then the states  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_T)$  is realized. In addition to the calibration errors as defined in Section 2.2, the algorithm is evaluated by the performance for omniprediction as in Definition 6, where the set of reference predictors  $\mathcal{B}$  is the set of predictors with low DTC to  $Q$ .

## 3 Smoothed Predictions for the Batch Setting

In this section, we will focus on post-processing in the batch setting where  $q$  is stochastically generated. Given a prediction  $q \sim Q$ , our privacy-based post-processing algorithm simply adds noise to  $q$ . We write the resulting predictor as  $P$ , with randomness from both  $Q$  and the privacy-based algorithm  $\mathcal{M}$ . Note that in the batch setting where predictions and states are stochastically drawn, the privacy-based post-processing algorithm optimizes for the expected error, where the expectation is taken with randomness from both the prediction, the state, and the post-processing algorithm.

- **Privacy-Based Post-Processing Algorithm  $\mathcal{M}$**
- **Input:** prediction  $q \sim Q$ , parameter  $\epsilon$  such that  $\text{DTC}(Q) \leq \epsilon$ , DP mechanism  $\mathcal{M}$ .
- **Output:** Prediction  $p \sim \mathcal{M}(q)$

Theorem 18 characterizes the decision loss of  $P$  with regard to all proper scoring rules and all predictors that are  $\epsilon$  close to  $Q$ .

---

<sup>2</sup> The algorithm in our paper only depends on  $q_t$ . This dependence on history only reinforces the definition.

► **Theorem 18.** *Suppose mechanism  $\mathcal{M}$  is  $(\gamma, \delta)$ -differentially private, then the output predictor  $P$  has at most  $C$  decision loss with regard to all proper scoring rules  $\mathcal{S}$  and the set of calibrated predictors  $\mathcal{B}$  such that any  $B \in \mathcal{B}$  is  $\epsilon$ -close to  $Q$ , i.e.  $\text{DIST}(Q, B) \leq \epsilon$ . The bound  $C$  is the following*

$$C \leq 2 \max_{q \in [0,1]} \mathbf{E} [|\mathcal{M}(q) - q|] + 4(1 - e^{-\gamma\epsilon} + \delta).$$

Moreover, ECE of  $P$  has the same bound.

We prove Theorem 18 following the idea of the Follow-The-Perturbed-Leader Algorithm [21]. We apply the same privacy-based post-processing algorithm  $\mathcal{M}$  to any calibrated predictor  $B$  that is  $\epsilon$  close to  $Q$ , which constructs a hypothetical predictor  $R$  as an intermediate connecting  $B$  and the post-processed predictor  $P = \mathcal{M}(Q)$ . Theorem 18 follows from combining Lemma 19 and Lemma 20, where Lemma 19 bounds the decision loss from  $B$  to  $R$ , and Lemma 20 characterizes the decision loss from  $R$  to  $P$  via DP mechanism  $\mathcal{M}$ .

► **Lemma 19.** *For any calibrated predictor  $B$ , we write  $R$  as the resulting predictor with the post-privacy-based processing algorithm  $\mathcal{M}$  applied to  $B$ . For any bounded proper scoring rule  $S(\cdot, \cdot) \in [0, 1]$ , the loss of  $R$  is bounded,*

$$DL(R) \leq 2 \max_{q \in [0,1]} \mathbf{E} [|\mathcal{M}(q) - q|].$$

The same bound holds for ECE.

$$ECE(R) \leq \max_{q \in [0,1]} \mathbf{E} [|\mathcal{M}(q) - q|].$$

► **Lemma 20.** *Suppose mechanism  $\mathcal{M}$  satisfies  $(\gamma, \delta)$ -differential privacy. We write  $R$  as the resulting predictor with the privacy-based post-processing algorithm applied to calibrated predictor  $B$  with  $\text{DIST}(Q, B) \leq \epsilon$ . The decision loss from  $R$  to  $P$  is bounded by*

$$\mathbf{E}_{(p,\theta) \sim D_{P,\Theta}} [S(p, \theta)] \geq \mathbf{E}_{(r,\theta) \sim D_{R,\Theta}} [S(r, \theta)] - 4(1 - e^{-\gamma\epsilon} + \delta).$$

A similar bound holds for ECE:

$$ECE(P) \leq ECE(R) + 4(1 - e^{-\gamma\epsilon} + \delta).$$

Lemma 21 shows the guarantee obtainable from some choices of the differentially private mechanism by adding noise  $D_\epsilon$ . We construct the noise by truncating the distribution with unbounded support into the feasible range of predictions. The parameters of  $(\gamma, \delta)$  are standard for Laplace and Gaussian noise [8].

► **Lemma 21.** *We consider two truncated noises that induce differential privacy.*

**Truncated Laplace** *Noise variable  $X$  from a truncated Laplace distribution with parameters  $(0, -\frac{1}{\ln \tau})$  is  $(-2 \ln \tau, 0)$ -differentially private. The expectation of the bias induced by noise is bounded:  $\mathbf{E} [|X|] \leq -\frac{1}{\ln \tau} - \frac{\tau}{1-\tau}$ . Combining the bounds and taking  $\tau = \exp(-\sqrt{\frac{1}{2\epsilon}})$ , we have  $C = \Theta(\sqrt{\epsilon})$ , the decision loss of the predictor is bounded by  $C$ , and  $ECE \leq C$ .*

**Truncated Gaussian** *Consider noise variable  $X$  from a truncated Gaussian distribution  $\mathcal{N}(0, 2\epsilon \ln(\frac{1.25}{\sqrt{\epsilon}}))$ . The truncated noise has*

$$\mathbf{E} [|X|] \leq \sigma = \sqrt{2\epsilon \ln(\frac{1.25}{\sqrt{\epsilon}})},$$

## 16:10 Smooth Calibration and Decision Making

and is  $(\gamma, \delta)$ -differentially private with  $\delta = \sqrt{\epsilon}$  and  $1 - e^{-\gamma\epsilon} \leq 2\sqrt{\epsilon}$ . Combining the bounds and taking  $C = \Theta(\sqrt{\epsilon \ln(\frac{1}{\epsilon})})$ , the decision loss of the predictor is bounded by  $C$ , and  $ECE \leq C$ .<sup>3</sup>

Theorem 22 shows that, there exists a predictor with  $DTC = \epsilon$ , such that no post-processing algorithm can achieve a worst-case decision loss better than  $\frac{\sqrt{\epsilon}}{2}$ . Our guarantee of decision loss in Theorem 18 is asymptotically optimal.

► **Theorem 22** (Post-processing Lowerbound for Batch Decision Loss). *There exists a predictor  $Q$ , with  $DTC(Q) = \epsilon$  and a reference calibrated predictor  $B \in \arg \min_{B'} \text{DIST}(B', Q)$ , such that for any post-processing algorithm that depends on the sequence  $\mathbf{q}$  of predictions,  $f_{\mathbf{q}}(q) : [0, 1] \rightarrow \Delta([0, 1])$ ,  $f_{\mathbf{q}}(Q)$  suffers a  $\frac{\sqrt{\epsilon}}{2}$  decision loss from  $B$ , i.e.*

$$\forall f, \exists S(\cdot, \cdot) \in [0, 1], \quad \mathbf{E}_{f, (p, \theta) \sim D_{P, \Theta}} [S(f_{\mathbf{q}}(q), \theta)] \leq \mathbf{E}_{(b, \theta) \sim D_{B, \Theta}} [S(b, \theta)] - \frac{\sqrt{\epsilon}}{2}.$$

As the main idea of the proof of lowerbound, any post-processing algorithm that does not depend on the state achieves a score at most by outputting the Bayesian posterior of predictor  $Q$ . We construct a predictor  $Q$  with a calibrated reference predictor  $B$  that is more informative than  $Q$ . By definition of  $DTC$  that specifies a coupling between  $B$ ,  $Q$ , and the state  $\theta$ , a reference calibrated predictor  $B$  may correlate with the state  $\theta$  when conditioned on  $Q$ . Thus, for this predictor  $Q$  and any post-processing algorithm  $f$ ,  $f(Q)$  achieves a lower score than  $B$ .

If the post-processing algorithm is a function of only the prediction  $Q$  but not the prediction sequence  $\mathbf{q}$ , our bounds are asymptotically optimal.

► **Corollary 23.** *For any post-processing algorithm that depends only on the current prediction,  $f(q) : [0, 1] \rightarrow \Delta([0, 1])$ , there exists a predictor  $Q$  with  $DTC(Q) = \epsilon$  and a reference calibrated predictor  $B \in \arg \min_{B'} \text{DIST}(B', Q)$ , such that  $f(Q)$  has*

$$ECE(Q) = \Theta(\sqrt{\epsilon}) \quad \text{and} \quad CDL(Q) = \Theta(\sqrt{\epsilon}).$$

Corollary 23 is a corollary from Theorem 27 which we will introduce later.

## 4 Smoothed Predictions for the Online Setting

To achieve guarantees for the online setting where the predictions  $\mathbf{q}$  and the states  $\theta$  are adversarially selected, the algorithm outputs are discretized for the empirical distribution to be meaningful. We prove empirical guarantees of the post-processing algorithm.

- **Input:** predictions  $q_t$ , parameter  $\epsilon$  such that  $DTC(\mathbf{q}, \theta) \leq \epsilon$ , DP mechanism  $\mathcal{M}$ .
- Discretize the space of predictions into  $T^{\frac{1}{3}}$  prediction values in  $\{\frac{i}{T^{\frac{1}{3}}} \mid i \in [T]\}$ .
- Draw  $p' \sim \mathcal{M}(q)$ .
- Find  $i$  such that  $p' \in [\frac{i}{T^{\frac{1}{3}}}, \frac{i+1}{T^{\frac{1}{3}}}]$ .
- **Output:**  $p = \frac{i}{T^{\frac{1}{3}}}$ .

<sup>3</sup> Section A.4 shows an improved  $O(\sqrt{\epsilon})$  bound for truncated Gaussian noise without the log factor. Note that we obtain Lemma 21 by bounding the TV-distance between the DP-mechanism output of adjacent predictions. Our improved bound directly analyzes this TV-distance rather than using the  $(\gamma, \delta)$  parameters of differential privacy.

By Theorem 18, the online privacy-based post-processing algorithm achieves the same bound for ECE up to a discretization error.

► **Theorem 24.** *Suppose mechanism  $\mathcal{M}$  is  $(\gamma, \delta)$ -differentially private. The output predictor  $\mathbf{p}$  satisfies*

$$\mathbf{E} [ECE(\mathbf{p}; \boldsymbol{\theta})] \leq \max_{q \in [0,1]} \mathbf{E} [|\mathcal{M}(q) - q|] + 4(1 - e^{-\gamma\epsilon} + \delta) + 2T^{-\frac{1}{3}}.$$

By Lemma 13, the same bound holds for CDL

► **Corollary 25.** *Suppose mechanism  $\mathcal{M}$  is  $(\gamma, \delta)$ -differentially private. The output predictor  $\mathbf{p}$  satisfies*

$$\mathbf{E} [CDL(\mathbf{p}; \boldsymbol{\theta})] \leq 2 \max_{q \in [0,1]} \mathbf{E} [|\mathcal{M}(q) - q|] + 8(1 - e^{-\gamma\epsilon} + \delta) + 2T^{-\frac{1}{3}}.$$

By Lemma 21, we obtain the guarantees for ECE and CDL in the online setting.

► **Lemma 26.** *With truncated Laplace noise, the privacy-based post-processing algorithm for online calibration achieves  $CDL \leq 2ECE = O(\sqrt{\epsilon}) + 2T^{-\frac{1}{3}}$ . With truncated Gaussian noise, the privacy-based post-processing algorithm achieves  $CDL \leq 2ECE = O(\sqrt{\epsilon \ln \frac{1}{\epsilon}}) + 2T^{-\frac{1}{3}}$ .*

Arunachaleswaran et al. [1] provides an online DTC minimization algorithm that achieves  $DTC = O(\frac{1}{\sqrt{T}})$ . Plugging into Lemma 26, the post-processing algorithm achieves  $ECE = O(T^{-\frac{1}{4}})$  with truncated Laplace noise and  $ECE = O(T^{-\frac{1}{4}} \ln T)$  with truncated Gaussian noise.

Theorem 27 shows that there exist two sequences of predictions  $\mathbf{q}, \mathbf{q}'$  and corresponding state realizations, such that both sequence has  $DTC = \epsilon$ . However, no post-processing algorithm can guarantee  $ECE < \Theta(\sqrt{\epsilon})$  or  $CDL < \Theta(\sqrt{\epsilon})$  for both sequences. Theorem 27 shows the online post-processing algorithm is asymptotically optimal for ECE as well as for CDL.

► **Theorem 27 (Post-processing Lowerbound for Online ECE).** *For any post-processing algorithm  $f = (f_1, \dots, f_T)$  where  $f_t$  depends on the prediction history  $(q_1, \dots, q_t)$  and  $(p_1, \dots, p_{t-1})$  before round  $t$ , there exists two sequences of predictions  $\mathbf{q}$  and  $\mathbf{q}'$  with states  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}'$ , respectively, both satisfying  $DTC(\mathbf{q}) = DTC(\mathbf{q}') = \epsilon$ , such that*

$$\max \{ \mathbf{E} [ECE(\mathbf{p}; \boldsymbol{\theta})], \mathbf{E} [ECE(\mathbf{p}'; \boldsymbol{\theta}')] \} \geq \frac{1}{8}\sqrt{\epsilon} + \frac{1}{2}\epsilon = \Theta(\sqrt{\epsilon}),$$

where we write  $\mathbf{p}, \mathbf{p}'$  as the output of the post-processing algorithm  $f$  on  $\mathbf{q}, \mathbf{q}'$ , respectively.

Moreover, the same lowerbound holds for CDL.

The lowerbound for CDL is perhaps surprising because Hu and Wu [19] shows a  $\tilde{O}(\frac{1}{\sqrt{T}})$  optimal bound for CDL, indicating ECE overestimates CDL when there exists a  $\omega(\frac{1}{\sqrt{T}})$  lowerbound for ECE [26]. We expected the same observation for the post-processing bound, which turns out not to be true. Considering the  $\epsilon = \Omega(T^{-\frac{2}{3}})$  lowerbound for DTC [27], the post-processing bound of  $O(\sqrt{\epsilon}) + 2T^{-\frac{1}{3}}$  is asymptotically optimal.

As an immediate corollary of our proof, even if the decision-makers are allowed to use different post-processing algorithms such as the differentially private exponential mechanism [24], there exists a worst-case decision-maker with a swap regret of  $\Theta(\sqrt{\epsilon})$ .

► **Corollary 28.** *There exists one decision-maker with proper scoring rule  $S$  such that for any post-processing algorithm  $f = (f_1, \dots, f_T)$  where  $f_t$  depends on the prediction history  $(q_1, \dots, q_t)$  and  $(p_1, \dots, p_{t-1})$  before round  $t$ , there exists two sequences of predictions  $\mathbf{q}$  and  $\mathbf{q}'$  with states  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}'$ , respectively, both satisfying  $DTC(\mathbf{q}) = DTC(\mathbf{q}') = \epsilon$ , such that*

$$\max \{ \mathbf{E} [\text{SWAP}_S(f(\mathbf{p}); \boldsymbol{\theta})], \mathbf{E} [\text{SWAP}_S(f(\mathbf{p}'); \boldsymbol{\theta}') ] \} \geq \frac{1}{8} \sqrt{\epsilon} + \frac{1}{2} \epsilon = \Theta(\sqrt{\epsilon}),$$

## 5 Discussion

Our lowerbound presents a gap in post-processing a predictor with a low distance to calibration from directly optimizing for calibration errors related to decision-making. However, in the examples we present, the conditional empirical frequencies are discontinuous in the prediction space, which does not match the discussion of machine learning predictors not distinguishing between small perturbations. One follow-up question is, are there properties of the predictor that, combined with a low distance to calibration, guarantee the predictor trustworthy for decision-making after post-processing?

Błasiok et al. [3] provides an answer to the question above. When the bias  $\hat{q} - q$  is 1-Lipschitz continuous in the prediction  $q$ , it follows that

$$\text{ECE}(Q) \leq O(\sqrt{DTC(Q)}),$$

and no post-processing algorithm is needed. This result, however, suggests the same problem as suggested by our lowerbound, that a given predictor with low DTC achieves a non-optimal ECE or CDL compared to optimizing for ECE or CDL directly in the online setting. Thus, it remains a question whether there exists a property of a predictor with low distance to calibration that guarantees an optimal ECE or CDL from post-processing.

---

## References

- 1 Eshwar Ram Arunachaleswaran, Natalie Collina, Aaron Roth, and Mirah Shi. An elementary predictor obtaining distance to calibration. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1366–1370. SIAM, 2025. doi:10.1137/1.9781611978322.41.
- 2 Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1727–1740, 2023.
- 3 Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 72071–72095. Curran Associates, Inc., 2023. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/e4165c96702bac5f4962b70f3cf2f136-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/e4165c96702bac5f4962b70f3cf2f136-Paper-Conference.pdf).
- 4 Jarosław Błasiok and Preetum Nakkiran. Smooth ece: Principled reliability diagrams via kernel smoothing. In *ICLR*, 2024. URL: <https://openreview.net/forum?id=XwiA1nDahv>.
- 5 Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. From external to swap regret 2.0: An efficient reduction and oblivious adversary for large action spaces. *arXiv preprint arXiv:2310.19786*, 2023. doi:10.48550/arXiv.2310.19786.
- 6 Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, Noah Golowich, Robert Kleinberg, and Princewill Okoroafor. Improved bounds for calibration via stronger sign preservation games. *arXiv preprint arXiv:2406.13668*, 2024. doi:10.48550/arXiv.2406.13668.
- 7 Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, Noah Golowich, Robert Kleinberg, and Princewill Okoroafor. Breaking the  $t^{2/3}$  barrier for sequential calibration, 2025.

- 8 Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi:10.1561/0400000042.
- 9 Dean P Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. *Games and Economic Behavior*, 109:271–293, 2018. doi:10.1016/J.GEB.2017.12.022.
- 10 Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40–55, 1997.
- 11 Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.
- 12 Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2725–2792, 2024. doi:10.1137/1.9781611977912.98.
- 13 Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:21, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.ITCS.2022.79.
- 14 Parikshit Gopalan, Michael Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 39936–39956. Curran Associates, Inc., 2023. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/7d693203215325902ff9dbdd067a50ac-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/7d693203215325902ff9dbdd067a50ac-Paper-Conference.pdf).
- 15 Parikshit Gopalan, Princewill Okoroafor, Prasad Raghavendra, Abhishek Shetty, and Mihir Singhal. Omnipredictors for regression and the approximate rank of convex functions. *arXiv preprint arXiv:2401.14645*, 2024. doi:10.48550/arXiv.2401.14645.
- 16 Nika Haghtalab, Mingda Qiao, Kunhe Yang, and Eric Zhao. Truthfulness of calibration measures. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- 17 Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 10–15 July 2018. URL: <https://proceedings.mlr.press/v80/hebert-johnson18a.html>.
- 18 Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 13497–13527. PMLR, 23–29 July 2023. URL: <https://proceedings.mlr.press/v202/hu23b.html>.
- 19 Lunjia Hu and Yifan Wu. Predict to minimize swap regret for all payoff-bounded tasks (calibration error for decision making). *65th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2024.
- 20 Sham M. Kakade and Dean P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008. Learning Theory 2004. doi:10.1016/j.jcss.2007.04.017.
- 21 Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005. doi:10.1016/J.JCSS.2004.10.016.
- 22 Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 247–254, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3306618.3314287.

- 23 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. URL: <https://proceedings.mlr.press/v195/kleinberg23a.html>.
- 24 Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007. doi:10.1109/FOCS.2007.41.
- 25 Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional unbiased prediction for sequential decision making. In *OPT 2023: Optimization for Machine Learning*, 2023. URL: <https://openreview.net/forum?id=P4j4145NUq>.
- 26 Mingda Qiao and Gregory Valiant. Stronger calibration lower bounds via sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, pages 456–466, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3406325.3451050.
- 27 Mingda Qiao and Letian Zheng. On the distance from calibration in sequential prediction. *arXiv preprint arXiv:2402.07458*, 2024. doi:10.48550/arXiv.2402.07458.

## A Missing Proof in Section 3

### A.1 Proof of Lemma 19

**Proof of Lemma 19.** Since  $S$  is bounded by  $[0, 1]$ , we know for any fixed  $b$ ,

$$\mathbf{E}_{\theta \sim b} [S(b, \theta) - S(r, \theta)] \leq 2|b - r|.$$

Thus,

$$\mathbf{E}_{b,r} [\mathbf{E}_{\theta \sim b} [S(b, \theta) - S(r, \theta)]] \leq 2\mathbf{E}_{b,r} [|b - r|],$$

which proves the argument for decision loss.

Now we prove Lemma 19 for ECE. We define  $Y(b_0) = \mathcal{M}(b_0) - b_0$ . The joint probability density function of state  $\theta$  and prediction value  $r$  can be expressed as

$$\begin{aligned} \Pr[\theta = 1, R = r] &= \int_0^1 \Pr[B = b] \cdot \Pr[\theta = 1, \mathcal{M}(b) = r | B = b] db \\ &= \int_0^1 \Pr[B = b] \cdot \Pr[\mathcal{M}(b) = r] \cdot \Pr[\theta = 1 | B = b] db \\ &= \int_0^1 \Pr[B = b] \cdot \Pr[\mathcal{M}(b) = r] \cdot b \, db. \end{aligned} \tag{2}$$

Equation (2) is derived given that  $B$  is a calibrated predictor.

According to the definition of ECE,

$$\begin{aligned} \text{ECE}(R) &= \mathbf{E}_R [|r - \Pr[\theta = 1 | R = r]|] \\ &= \int_0^1 \Pr[R = r] \cdot |r - \Pr[\theta = 1 | R = r]| dr \\ &= \int_0^1 |r \Pr[R = r] - \Pr[\theta = 1, R = r]| dr \\ &= \int_0^1 \left| \int_0^1 \Pr[\mathcal{M}(b) = r] \cdot \Pr[B = b] \cdot (r - b) \, db \right| dr \\ &\leq \int_0^1 \int_0^1 \Pr[\mathcal{M}(b) = r] \cdot \Pr[B = b] \cdot |r - b| \, db dr \\ &= \mathbf{E} [|\mathcal{M}(b) - b|]. \end{aligned} \quad \blacktriangleleft$$

## A.2 Proof of Lemma 20

**Proof of Lemma 20.** We prove the lemma by Lemma 29, bounding the TV-distance between  $\mathcal{M}(B)$  and  $\mathcal{M}(Q)$ . Combining with Lemma 30, we prove Lemma 20.  $\blacktriangleleft$

**► Lemma 29.** *We write  $R$  as the resulting predictor with post-processing algorithm applied to calibrated predictor  $B$  with  $\text{DIST}(Q, B) \leq \epsilon$ . The decision loss from  $R$  to  $P$  is bounded by*

$$\mathbf{E}_{(p,\theta) \sim D_{P,\Theta}} [S(p, \theta)] \geq \mathbf{E}_{(r,\theta) \sim D_{R,\Theta}} [S(r, \theta)] - 4\mathbf{E}_{(b,q) \sim D_{B,Q}} [d_{TV}(\mathcal{M}(b), \mathcal{M}(q))].$$

Note that the TV distance quantifies the distance between  $\mathcal{M}(b)$  and  $\mathcal{M}(q)$ .

A similar bound holds for ECE:

$$\text{ECE}(P) \leq \text{ECE}(R) + 4\mathbf{E}_{(b,q) \sim D_{B,Q}} [d_{TV}(\mathcal{M}(b), \mathcal{M}(q))].$$

Lemma 29 follows from the fact that the scoring rule  $S$  is bounded in  $[0, 1]$ .

**Proof of Lemma 29.** Since the scoring rule  $S$  is bounded in  $[0, 1]$ , we know for any fixed  $b$  and  $q$ ,

$$\begin{aligned} & \mathbf{E}_{r \sim \mathcal{M}(b), \theta \sim b} [S(r, \theta)] - \mathbf{E}_{p \sim \mathcal{M}(q), \theta \sim b} [S(p, \theta)] \\ &= \int_0^1 (\Pr[\mathcal{M}(b) = p] - \Pr[\mathcal{M}(q) = p]) \mathbf{E}_{\theta \sim b} [S(p, \theta)] dp \\ &\leq 4d_{TV}(\mathcal{M}(b), \mathcal{M}(q)). \end{aligned}$$

Thus Lemma 29 for decision loss from  $R$  to  $P$  holds.

Now we prove Lemma 29 for ECE by dividing it into three parts, the first part is

$$\begin{aligned} & \int_0^1 p |\Pr[P = p] - \Pr[R = p]| dp \\ &= \int_0^1 p \left| \int_0^1 \int_0^1 \Pr[B = b, Q = q] (\Pr[\mathcal{M}(q) = p] - \Pr[\mathcal{M}(b) = p]) dbdq \right| dp \\ &\leq \int_0^1 \int_0^1 \Pr[B = b, Q = q] \int_0^1 |\Pr[\mathcal{M}(b) = p] - \Pr[\mathcal{M}(q) = p]| dp dbdq \\ &= 2\mathbf{E}_{(b,q) \sim D_{B,Q}} [d_{TV}(\mathcal{M}(b), \mathcal{M}(q))]. \end{aligned} \tag{3}$$

The distance between joint distribution  $D_{R,\Theta}$  and  $D_{P,\Theta}$  is

$$\begin{aligned} & \int_0^1 |\Pr[R = p] \Pr[\theta = 1 | R = p] - \Pr[P = p] \Pr[\theta = 1 | P = p]| dp \\ &= \int_0^1 |\Pr[\theta = 1, R = p] - \Pr[\theta = 1, P = p]| dp \\ &\leq \int_0^1 \int_0^1 \Pr[B = b, Q = q] \int_0^1 |\Pr[\mathcal{M}(b) = p] - \Pr[\mathcal{M}(q) = p]| dp dbdq \\ &= 2\mathbf{E}_{(b,q) \sim D_{B,Q}} [d_{TV}(\mathcal{M}(b), \mathcal{M}(q))]. \end{aligned} \tag{4}$$

## 16:16 Smooth Calibration and Decision Making

Combine (3) and (4),

$$\begin{aligned}
\text{ECE}(P) &= \int_0^1 \Pr[P = p] |p - \Pr[\theta = 1 \mid P = p]| dp \\
&\leq \int_0^1 |\Pr[R = p] (p - \Pr[\theta = 1 \mid R = p])| dp \\
&\quad + \int_0^1 p |\Pr[P = p] - \Pr[R = p]| dp \\
&\quad + \int_0^1 |\Pr[R = p] \Pr[\theta = 1 \mid R = p] - \Pr[P = p] \Pr[\theta = 1 \mid P = p]| dp \\
&\leq \text{ECE}(R) + 4\mathbf{E}_{(b,q) \sim D_{B,Q}} [d_{\text{TV}}(\mathcal{M}(b), \mathcal{M}(q))]. \quad \blacktriangleleft
\end{aligned}$$

► **Lemma 30.** *Given noise  $X, Y$  for  $(\gamma, \delta)$ -differential privacy,  $X$  and  $Y$  are drawn from the same distribution,*

$$\mathbf{E}_{(b,q) \sim D_{B,Q}} [d_{\text{TV}}(\mathcal{M}(b), \mathcal{M}(q))] \leq 1 - e^{-\gamma\epsilon} + \delta.$$

**Proof of Lemma 30.** For any pair of fixed  $(q, b)$ , consider the set of prediction values  $V = \{p \mid \Pr[\mathcal{M}(b) = p] - \Pr[\mathcal{M}(q) = p] \leq 0\}$ .

By Definition 1 of differential privacy,

$$\begin{aligned}
\Pr[\mathcal{M}(b) = p] - \Pr[\mathcal{M}(q) = p] &\geq e^{-\gamma|b-q|} (\Pr[\mathcal{M}(q) = p] - \delta) - \Pr[\mathcal{M}(q) = p] \\
&= (e^{-\gamma|b-q|} - 1) \Pr[\mathcal{M}(q) = p] - \delta e^{-\gamma|b-q|}.
\end{aligned}$$

Calculate  $d_{\text{TV}}(\mathcal{M}(b), q + X)$  using prediction values in  $V$ :

$$\begin{aligned}
d_{\text{TV}}(\mathcal{M}(b), \mathcal{M}(q)) &= \int_V |\Pr[\mathcal{M}(b) = p] - \Pr[\mathcal{M}(q) = p]| dp \\
&\leq \int_V \left[ (1 - e^{-\gamma|b-q|}) \Pr[\mathcal{M}(q) = p] + \delta e^{-\gamma|b-q|} \right] dp \\
&\leq 1 - e^{-\gamma|b-q|} + \delta e^{-\gamma|b-q|}.
\end{aligned}$$

Take the expectation with respect to  $(b, p)$  and get

$$\begin{aligned}
\mathbf{E}_{(b,q) \sim D_{B,Q}} [d_{\text{TV}}(\mathcal{M}(b), \mathcal{M}(q))] &\leq \mathbf{E}_{(b,q) \sim D_{B,Q}} \left[ 1 - (1 - \delta) e^{-\gamma|b-q|} \right] \\
&\leq 1 - (1 - \delta) e^{-\gamma\epsilon} \\
&\leq 1 - e^{-\gamma\epsilon} + \delta.
\end{aligned}$$

The second inequality follows from Jensen's inequality, give that  $1 - \delta \geq 0$ , function  $e^{-\gamma x}$  is convex and  $\mathbf{E}_{(b,q) \sim D_{B,Q}} [|b - p|] = \epsilon$ . ◀

Similarly combining Lemma 20 and Lemma 19, post-processed predictor  $P$  is calibrated in ECE.

### A.3 Proof of Lemma 21

#### A.3.1 Truncated Laplace Noise

**Proof of Lemma 21, Truncated Laplace.** For  $\forall q, p \in [0, 1]$  and differentially private mechanism  $\mathcal{M}$  as adding noise from the truncated Laplace distribution,

$$\Pr[\mathcal{M}(q) = p] = \frac{-\ln \tau}{2 - \tau^q - \tau^{1-q}} \cdot \tau^{|p-q|}.$$

$$\frac{\Pr[\mathcal{M}(q) = p]}{\Pr[\mathcal{M}(q') = p]} = \tau^{|p-q| - |p-q'|} \cdot \frac{2 - \tau^{q'} - \tau^{1-q'}}{2 - \tau^q - \tau^{1-q}}.$$

Since  $|p - q| - |p - q'| \geq -|q - q'|$ ,

$$\tau^{|p-q| - |p-q'|} \leq \tau^{-|q-q'|}.$$

The following steps will show

$$\frac{2 - \tau^{q'} - \tau^{1-q'}}{2 - \tau^q - \tau^{1-q}} \leq \tau^{-|q-q'|}.$$

Case 1:  $q' \leq q$ .

$$\begin{aligned} \frac{2 - \tau^{q'} - \tau^{1-q'}}{2 - \tau^q - \tau^{1-q}} &\leq \tau^{-|q-q'|} \\ \Leftrightarrow -\tau^{q+1} \left(\tau^{-q'}\right)^2 + 2\tau^q \cdot \tau^{q'} + \tau^{1-q} &\leq 2. \end{aligned}$$

Since  $\tau^{-q'} \in [1, \tau^{-q}]$ ,  $-\tau^{q+1} \left(\tau^{-q'}\right)^2 + 2\tau^q \cdot \tau^{q'} + \tau^{1-q}$  achieves its maximum value at  $\tau^{-q}$ , and the maximum value is 2.

Case 2:  $q' \geq q$ .

$$\begin{aligned} \frac{2 - \tau^{q'} - \tau^{1-q'}}{2 - \tau^q - \tau^{1-q}} &\leq \tau^{-|q-q'|} \\ \Leftrightarrow -\tau^{-q} \left(\tau^{q'}\right)^2 + 2\tau^{-q} \cdot \tau^{q'} + \tau^q &\leq 2. \end{aligned}$$

Since  $\tau^{-q'} \in [\tau, \tau^q]$ ,  $-\tau^{-q} \left(\tau^{q'}\right)^2 + 2\tau^{-q} \cdot \tau^{q'} + \tau^q$  achieves its maximum value at  $\tau^q$ , and the maximum value is 2.

Therefore,

$$\Pr[\mathcal{M}(q) = p] \leq \tau^{-2|q-q'|} \Pr[\mathcal{M}(q') = p].$$

For any subset  $\mathcal{I} \subseteq [0, 1]$  of predictions,

$$\Pr[\mathcal{M}(q) \in \mathcal{I}] \leq \tau^{-2|q-q'|} \cdot \Pr[\mathcal{M}(q') \in \mathcal{I}]. \quad \blacktriangleleft$$

### A.3.2 Truncated Gaussian Noise

**Proof of Lemma 21, Truncated Gaussian.** The choice of parameters is adopted from Dwork et al. [8]. We write the proof here for reference. The proof has two main steps. First, we show that the Gaussian distribution  $Y \sim \mathcal{N}\left(0, 2\epsilon \ln\left(\frac{1.25}{\sqrt{\epsilon}}\right)\right)$  is  $(\gamma_0, \delta)$ -differentially private with  $\delta = \sqrt{\epsilon}$  and  $1 - e^{-\gamma_0\epsilon} \leq \sqrt{\epsilon}$ . Then we show the probability that Gaussian is truncated is bounded by  $1 - \exp\left(-\frac{1}{4\sqrt{\epsilon}}\right)$ , implying

$$\frac{\Pr[X = p]}{\Pr[Y = p]} \leq \frac{1}{1 - \exp\left(-\frac{1}{4\sqrt{\epsilon}}\right)}.$$

By Definition 1, the truncated distribution has  $\delta = \sqrt{\epsilon}$  and  $1 - e^{-\gamma\epsilon} \leq 1 - e^{-\gamma_0\epsilon}(1 - \exp(-\frac{1}{4\sqrt{\epsilon}})) \leq 2\sqrt{\epsilon}$ .

## 16:18 Smooth Calibration and Decision Making

Now we show Gaussian distribution  $Y \sim \mathcal{N}\left(0, 2\epsilon \ln\left(\frac{1.25}{\sqrt{\epsilon}}\right)\right)$  is differentially private. Notice that for Definition 1, it suffices to show

$$\Pr_{p \sim q+Y} \left[ \frac{\Pr_Y[q+Y=p]}{\Pr_Y[q'+Y=p]} \geq e^{\gamma_0|q-q'|} \right] \leq \delta.$$

Define  $L(p) = \frac{\Pr_Y[q+Y=p]}{\Pr_Y[q'+Y=p]}$ . We know

$$\ln[L(p)] = \frac{-(p-q)^2 + (p-q')^2}{2\sigma^2} = \frac{(q-q')^2 + 2(p-q) \cdot (q'-q')}{2\sigma^2},$$

where  $(p-q)$  is the Gaussian  $\mathcal{N}(0, \sigma^2)$ . Applying the tail bound for Gaussian distribution with  $\gamma_1 = \gamma_0|q-q'|$

$$\Pr[\ln[L(p)] \geq \gamma_1] \leq \exp\left(-\frac{(\gamma_1\sigma^2 - \frac{1}{2}(q-q')^2)^2}{(q'-q)^2\sigma^2}\right)$$

$$\text{For } \sigma = \sqrt{2\epsilon \ln\left(\frac{1.25}{\delta}\right)} \geq \frac{\sqrt{2\ln\left(\frac{1.25}{\delta}\right)} \cdot |q-q'|}{\gamma_1}, \text{ we have } \Pr[\ln[L(p)] \leq \delta]. \quad \blacktriangleleft$$

### A.4 Improved Bound for Truncated Gaussian Noise

For any distribution  $D$  with probability density function  $f$ , define  $f^b(x)$  as the probability density function of truncated distribution of  $D$  on the interval  $[-b, 1-b]$  and  $f_b(x) = \frac{f(x)}{\int_{-b}^{1-b} f(x)dx}$ .

► **Lemma 31.** *Consider any distribution of noise with probability density function  $f(x)$  that is monotone on  $x \geq 0$  and  $x < 0$  respectively. Then for  $\forall b, q \in [0, 1]$ ,*

$$d_{TV}(\mathcal{M}(b), \mathcal{M}(q)) \leq \max\{f^b(x), f^q(x)\} \cdot |q-b|.$$

**Proof of Lemma 31.** Fix  $b$  and  $q$ , without loss of generality, assume that  $b \leq q$ . There exists  $t \in [b, q]$  that  $f^b(t-b) = f^q(t-q)$ .

Represent the probability of  $\mathcal{M}(b) \in [0, t]$  by  $S = \int_{-b}^{t-b} f^b(x)dx$ , so  $d_{TV}(\mathcal{M}(b), \mathcal{M}(q)) = S - \int_{-q}^{t-q} f^q(x)dx$ .

When  $t \geq q-b$ , represent the probability of  $\mathcal{M}(b) \in [t+b-q, t]$  by  $S_1 = \int_{t-q}^{t-b} f^b(x)dx = S - \int_{-b}^{t-q} f^b(x)dx$ . The aim is to show  $d_{TV}(\mathcal{M}(b), \mathcal{M}(q)) \leq S_1$ .

(i) If  $\int_{-b}^{1-b} f(x)dx \geq \int_{-q}^{1-q} f(x)dx$ , then

$$\begin{aligned} d_{TV}(\mathcal{M}(b), \mathcal{M}(q)) &\leq S_1 \\ \Leftrightarrow \int_{-b}^{t-q} f^b(x)dx &\leq \int_{-q}^{t-q} f^q(x)dx \\ \Leftrightarrow \int_{-b}^{t-q} (f^b(x) - f^q(x)) dx &\leq \int_{-q}^{-b} f^q(x)dx. \end{aligned}$$

(ii) If  $\int_{-b}^{1-b} f(x)dx < \int_{-q}^{1-q} f(x)dx$ , then  $\int_{t-q}^0 f_b(x)dx > \int_{t-q}^0 f_q(x)dx$ . Since

$$\int_{-b}^0 f_b(x)dx = \frac{\int_{-b}^0 f(x)dx}{\int_{-b}^{1-b} f(x)dx} = \frac{1}{1 + \frac{\int_0^{1-b} f(x)dx}{\int_{-b}^0 f(x)dx}}$$

is an increasing function of  $b$ ,

$$\int_{-b}^0 f_b(x)dx \leq \int_{-q}^0 f_q(x)dx.$$

$$\begin{aligned} d_{\text{TV}}(\mathcal{M}(b), \mathcal{M}(q)) &\leq S_1 \\ \Leftrightarrow \int_{-b}^{t-q} f_b(x)dx &\leq \int_{-q}^{t-q} f_q(x)dx \\ \Leftrightarrow \int_{-b}^0 f_b(x)dx - \int_{t-q}^0 f_b(x)dx &\leq \int_{-q}^0 f_q(x)dx - \int_{t-q}^0 f_q(x)dx \end{aligned}$$

Therefore,

$$d_{\text{TV}}(\mathcal{M}(b), \mathcal{M}(q)) \leq S_1 \leq (q - b) \max\{f^b(x), f^q(x)\}.$$

When  $t < q - b$ ,

$$d_{\text{TV}}(\mathcal{M}(b), \mathcal{M}(q)) = S - \int_{-q}^{t-q} f^q(x)dx < S < (q - b) \max\{f^b(x), f^q(x)\}. \quad \blacktriangleleft$$

► **Lemma 32.** Consider adding the truncated noise from a Gaussian distribution  $\mathcal{N}(0, \sqrt{\epsilon})$  in the same way as Lemma 21, then for  $C = \Theta(\sqrt{\epsilon})$ , the predictor is  $C$ -omnipredictor with  $ECE \leq C$ .

**Proof.** The truncated noise has

$$\mathbf{E}[|X|] \leq \sigma = \sqrt{\epsilon}.$$

The maximum value of the truncated Gaussian distribution's probability density function is

$$\max_{q,p \in [0,1]} Pr_{p \sim q+X}[q+X = p] = \max_{q,p \in [0,1]} \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(p-q)^2}{2\sigma^2}\right)}{\int_{-q}^{1-q} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx} = \frac{\frac{1}{\sqrt{2\pi}\sigma}}{\int_0^1 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx}.$$

Since  $\exp\left(-\frac{x^2}{2\sigma^2}\right)$  is concave on  $[0, \sigma]$ ,  $\int_0^\sigma \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$  can be lower bounded by the area of a ladder:

$$\int_0^1 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \geq \frac{1}{2\sqrt{2\pi}\sigma} \left(1 + \exp\left(-\frac{1}{2}\right)\right) \cdot \sigma \geq \frac{1}{2\sqrt{2\pi}}.$$

By Lemma 31,

$$\mathbf{E}_{(b,q) \sim D_{B,Q}}[d_{\text{TV}}(\mathcal{M}(b), \mathcal{M}(q))] \leq \mathbf{E}_{(b,q) \sim D_{B,Q}}\left[\frac{2}{\sigma} |q - b|\right] = \frac{2\epsilon}{\sigma}.$$

Therefore, the parameter  $C$  of the predictor can be upper bounded by  $\sigma + \frac{8\epsilon}{\sigma} = \Theta(\sqrt{\epsilon})$ . ◀

## A.5 Proof of Theorem 22

**Proof of Theorem 22.** Fix a predictor  $Q$ , define predictor  $\tilde{Q}$  that predicts the Bayesian posterior of  $Q$ : for every prediction value  $q_i$ , when  $Q$  predicts  $q_i$ , let  $\tilde{Q}$  predict  $\hat{q}_i = Pr[\theta = 1 | q = q_i]$ . Post-process predictor  $Q$  by  $f$  and get predictor  $P$ .

## 16:20 Smooth Calibration and Decision Making

Fix a prediction value  $q_i$  and a proper scoring rule  $S$ , consider all predictions  $p \sim f(q_i)$ , according to the definition of proper scoring rules, the score achievable by  $f$  is upperbounded by  $\widehat{Q}$ :

$$\mathbf{E}_{p \sim f(q_i)} \left[ \mathbf{E}_{\theta \sim \widehat{q}_i} [S(p, \theta)] \right] \leq \mathbf{E}_{p \sim f(q_i)} \left[ \mathbf{E}_{\theta \sim \widehat{q}_i} [S(\widehat{q}_i, \theta)] \right] = \mathbf{E}_{\theta \sim \widehat{q}_i} [S(\widehat{q}_i, \theta)].$$

$$\begin{aligned} \mathbf{E}_{(p, \theta) \sim D_{P, \Theta}} [S(p, \theta)] &= \mathbf{E}_{q_i \sim q} \left[ \mathbf{E}_{p \sim f(q_i)} \left[ \mathbf{E}_{\theta \sim \widehat{q}_i} [S(p, \theta)] \right] \right] \\ &\leq \mathbf{E}_{q_i \sim q} \left[ \mathbf{E}_{\theta \sim \widehat{q}_i} [S(\widehat{q}_i, \theta)] \right] = \mathbf{E}_{(p, \theta) \sim D_{\widehat{Q}, \Theta}} [S(p, \theta)]. \end{aligned}$$

Consider the following predictor  $Q$  with  $\text{DTC}(Q) = \epsilon$ .

Case 1: With probability  $1 - \sqrt{\epsilon}$ , the distribution of predictions and states follows

$$(q, \widehat{q}) = \begin{cases} (\frac{1}{2} - \sqrt{\epsilon}, \frac{1}{2} - \sqrt{\epsilon}) & \text{w.p. } \frac{1}{2} \\ (\frac{1}{2} + \sqrt{\epsilon}, \frac{1}{2} + \sqrt{\epsilon}) & \text{w.p. } \frac{1}{2} \end{cases}$$

Case 2: With probability  $\sqrt{\epsilon}$ , the distribution of predictions and states follows

$$(q, \widehat{q}) = \begin{cases} (\frac{1}{2} - \sqrt{\epsilon}, 1) & \text{w.p. } \frac{1}{2} \\ (\frac{1}{2} + \sqrt{\epsilon}, 0) & \text{w.p. } \frac{1}{2} \end{cases}$$

Therefore the corresponding  $\widetilde{q}$  follows

$$(\widetilde{q}, q) = \begin{cases} (\frac{1}{2} - \frac{1}{2}\sqrt{\epsilon} + \epsilon, \frac{1}{2} - \sqrt{\epsilon}) & \text{w.p. } \frac{1}{2} \\ (\frac{1}{2} + \frac{1}{2}\sqrt{\epsilon} - \epsilon, \frac{1}{2} + \sqrt{\epsilon}) & \text{w.p. } \frac{1}{2} \end{cases}$$

Define a calibrated predictor  $B$ , when  $Q$  follows from Case 1, let  $B$  outputs the same prediction of  $Q$ . When  $Q$  follows from Case 2, let  $B$  always predicts  $\frac{1}{2}$ . Notice that

$$\text{DTC}(Q) \leq \text{DIST}(Q, B) = \epsilon,$$

to show  $\text{DTC}(Q) = \epsilon$ , use a linear program with infinite constraints to prove  $\text{DTC}(Q) \geq \epsilon$ . Notice that  $\mathcal{Q} = \{\frac{1}{2} - \sqrt{\epsilon}, \frac{1}{2} + \sqrt{\epsilon}\}$ . Let  $\rho$  denotes joint probability distribution function of  $(b, q, \theta) \in [0, 1] \times \mathcal{Q} \times \{0, 1\}$ . The following linear program is feasible and its optimal value equals  $\text{DTC}(Q)$ .

$$\begin{aligned} &\text{minimize} && \sum_{(b, q, \theta) \in [0, 1] \times \mathcal{Q} \times \{0, 1\}} |q - b| \rho(b, q, \theta) && (5) \\ &\text{s.t.} && \sum_{b \in [0, 1]} \rho(b, q, \theta) = \text{Pr}[q, \theta], && \text{for } \forall (q, \theta) \in \mathcal{Q} \times \{0, 1\}; \quad (r(q, \theta)) \\ &&& (1 - b) \sum_{q \in \mathcal{Q}} \rho(b, q, 1) - b \sum_{q \in \mathcal{Q}} \rho(b, q, 0) = 0, && \text{for } \forall b \in [0, 1]; \quad (s(b)) \\ &&& \rho(b, q, \theta) \geq 0, && \text{for } \forall (b, q, \theta) \in [0, 1] \times \mathcal{Q} \times \{0, 1\}. \end{aligned}$$

The objective of this linear program corresponds to  $\text{DTC}(Q)$ . The first constraint ensures that the joint distribution of  $(b, p, \theta)$  is consistent with the joint distribution of  $(q, \theta)$ . The second constraint ensures that predictor  $B$  is calibrated. This linear program is feasible, because

$$\rho(b, q, \theta) = \begin{cases} \text{Pr}[q, \theta] & \text{if } b = \theta \\ 0 & \text{else} \end{cases}$$

is a feasible solution of this linear program. The dual of the linear program (5) is

$$\begin{aligned} & \text{maximize} && \sum_{(q,\theta) \in \mathcal{Q} \times \{0,1\}} Pr [q, \theta] r(q, \theta) \\ & \text{s.t.} && r(q, \theta) \leq |b - q| + (\theta - b)s(b), \quad \text{for } \forall (b, q, \theta) \in [0, 1] \times \mathcal{Q} \times \{0, 1\}. \end{aligned} \quad (6)$$

If  $s(b) > 1$ , change  $s(b)$  to 1 still satisfy the constraints and the objective stays the same:

$$\begin{aligned} r(q, 0) &\leq |b - q| - bs(b) < |b - q| - b, \\ r(q, 1) &\leq |1 - q| \leq |b - q| + (1 - b). \end{aligned}$$

If  $s(b) < -1$ , change  $s(b)$  to  $-1$  still satisfy the constraints and the objective stays the same:

$$\begin{aligned} r(q, 0) &\leq q < |b - q| + b, \\ r(q, 1) &\leq |b - q| + (1 - b)s(q) \leq |b - q| - (1 - b). \end{aligned}$$

Therefore, the optimal solution of linear program (6) stays the same after adding the constraints:

$$-1 \leq s(b) \leq 1, \quad \text{for } \forall b \in [0, 1].$$

The optimal value of linear program (5) can be lower bounded by the objective of linear program (6):

$$\begin{aligned} & \sum_{(q,\theta) \in \mathcal{Q} \times \{0,1\}} Pr [q, \theta] r(q, \theta) \\ = & \sum_{(q,\theta) \in \mathcal{Q} \times \{0,1\}} r(q, \theta) \sum_{b \in [0,1]} \rho(b, q, \theta) + \sum_{b \in [0,1]} s(b) \sum_{(q,\theta) \in \mathcal{Q} \times \{0,1\}} (b - \theta) \rho(b, q, \theta) \end{aligned} \quad (7)$$

$$= \sum_{(q,\theta) \in \mathcal{Q} \times \{0,1\}} \sum_{b \in [0,1]} r(q, \theta) \rho(b, q, \theta) + \sum_{b \in [0,1]} \sum_{(q,\theta) \in \mathcal{Q} \times \{0,1\}} s(b) (b - \theta) \rho(b, q, \theta) \quad (8)$$

$$= \sum_{(b,q,\theta) \in [0,1] \times \mathcal{Q} \times \{0,1\}} [r(q, \theta) + (b - \theta)s(b)] \rho(b, q, \theta) \quad (9)$$

$$\leq \sum_{(b,q,\theta) \in [0,1] \times \mathcal{Q} \times \{0,1\}} |q - b| \rho(b, q, \theta).$$

(7)=(8) holds because  $\sum_{b \in [0,1]} \rho(b, q, \theta)$  is absolutely convergent, the distributive property of multiplication still holds. (8)=(9) holds because Equation (8) is absolutely convergent, the commutative property of addition still holds.

Let

$$s(b) = \begin{cases} \frac{2\sqrt{\epsilon}}{2\sqrt{\epsilon}+1} & \text{if } b < \frac{1}{2} \\ 0 & \text{if } b = \frac{1}{2} \\ -\frac{2\sqrt{\epsilon}}{2\sqrt{\epsilon}+1} & \text{if } b > \frac{1}{2} \end{cases}$$

Then the constraints for the dual linear program (6) are

$$\begin{aligned} r\left(\frac{1}{2} - \sqrt{\epsilon}, 0\right) &\leq \min_{b \in [0,1]} \left\{ \left| b - \frac{1}{2} + \sqrt{\epsilon} \right| - bs(b) \right\} = \frac{-\sqrt{\epsilon}(1 - 2\sqrt{\epsilon})}{2\sqrt{\epsilon} + 1}, \\ r\left(\frac{1}{2} - \sqrt{\epsilon}, 1\right) &\leq \min_{b \in [0,1]} \left\{ \left| b - \frac{1}{2} + \sqrt{\epsilon} \right| + (1 - b)s(b) \right\} = \sqrt{\epsilon}, \\ r\left(\frac{1}{2} + \sqrt{\epsilon}, 0\right) &\leq \min_{b \in [0,1]} \left\{ \left| b - \frac{1}{2} - \sqrt{\epsilon} \right| - bs(b) \right\} = \sqrt{\epsilon}, \\ r\left(\frac{1}{2} + \sqrt{\epsilon}, 1\right) &\leq \min_{b \in [0,1]} \left\{ \left| b - \frac{1}{2} - \sqrt{\epsilon} \right| + (1 - b)s(b) \right\} = \frac{-\sqrt{\epsilon}(1 - 2\sqrt{\epsilon})}{2\sqrt{\epsilon} + 1}. \end{aligned}$$

## 16:22 Smooth Calibration and Decision Making

Take maximum values of all  $r(q, \theta)$  and get the optimal value of linear program (6) is no less than

$$\begin{aligned} & \frac{1}{2} \left( \frac{1}{2} + \frac{\sqrt{\epsilon}}{2} - \epsilon \right) \left[ r \left( \frac{1}{2} - \sqrt{\epsilon}, 0 \right) + r \left( \frac{1}{2} + \sqrt{\epsilon}, 1 \right) \right] \\ & + \frac{1}{2} \left( \frac{1}{2} - \frac{\sqrt{\epsilon}}{2} + \epsilon \right) \left[ r \left( \frac{1}{2} - \sqrt{\epsilon}, 1 \right) + r \left( \frac{1}{2} + \sqrt{\epsilon}, 0 \right) \right] = \epsilon. \end{aligned}$$

Therefore,  $\text{DTC}(Q) \geq \epsilon$  and thus  $\text{DTC}(Q) = \epsilon$ .

Consider the proper scoring rule

$$S(p, \theta) = \begin{cases} 1 - \theta & \text{if } p \leq \frac{1}{2} \\ \theta & \text{if } p > \frac{1}{2} \end{cases}$$

and calculate the expected payoff in decision making for predictor  $Q$  and  $B$ :

$$\mathbf{E}_{(p, \theta) \sim D_{Q, \Theta}} [S(p, \theta)] = \frac{1}{2} + \frac{1}{2} \sqrt{\epsilon} - \epsilon.$$

$$\mathbf{E}_{(b, \theta) \sim D_{B, \Theta}} [S(b, \theta)] = \frac{1}{2} + \sqrt{\epsilon} - \epsilon.$$

Therefore, for any post-processed algorithm  $f$ , there exists predictor  $Q$  and a reference calibrated predictor  $b$  such that  $\text{DTC}(Q) = \epsilon$  and

$$\text{DL}(f(Q); B) \geq \mathbf{E}_{(b, \theta) \sim D_{B, \Theta}} [S(b, \theta)] - \mathbf{E}_{(p, \theta) \sim D_{Q, \Theta}} [S(p, \theta)] = \frac{\sqrt{\epsilon}}{2}. \quad \blacktriangleleft$$

## B Missing Proof in Section 4

### B.1 Proof of Theorem 24

**Proof of Theorem 24.** We write  $n_i$  as the number of times that  $\frac{i}{T^{\frac{1}{3}}}$  is predicted. Clearly,  $\sum_{i \in [eT]} n_i = T$ . We also write  $p'_t$  as the output of post-processed predictor before discretization. Conditioning on a set of  $(n_i)_i$ , we know for each  $n_i$ :

$$\begin{aligned} & \mathbf{E} \left[ \left| \frac{i}{T^{\frac{1}{3}}} - \sum_t \mathbb{I} \left[ p_t = \frac{i}{T^{\frac{1}{3}}} \right] \frac{\theta_t}{n_i} \right| \right] \\ & \leq \mathbf{E} \left[ \left| \frac{i}{T^{\frac{1}{3}}} - \sum_t \mathbb{I} \left[ p_t = \frac{i}{T^{\frac{1}{3}}} \right] p'_t \right| \right] + \frac{1}{n_i} \mathbf{E} \left[ \left| \sum_t \mathbb{I} \left[ p_t = \frac{i}{T^{\frac{1}{3}}} \right] \theta_t - \sum_t \mathbb{I} \left[ p_t = \frac{i}{T^{\frac{1}{3}}} \right] p'_t \right| \right] \\ & \leq T^{-\frac{1}{3}} + \sqrt{\text{Var} \left[ \sum_t \mathbb{I} \left[ p_t = \frac{i}{T^{\frac{1}{3}}} \right] \frac{\theta_t}{n_i} \right]} \\ & \quad + \frac{1}{n_i} \mathbf{E} \left[ \sum_t \mathbb{I} \left[ p_t = \frac{i}{T^{\frac{1}{3}}} \right] \left| \Pr[\theta | p'_t] - \sum_t \mathbb{I} \left[ p_t = \frac{i}{T^{\frac{1}{3}}} \right] p'_t \right| \right], \end{aligned}$$

where  $\Pr[\theta | p'_t]$  is defined over the empirical distribution over  $T$  rounds with the noise of the algorithm. Summing over all prediction values, we know

$$\frac{1}{T} \sum_i \mathbf{E} \left[ \left| \frac{i}{T^{\frac{1}{3}}} - \sum_t \mathbb{I} \left[ p_t = \frac{i}{\epsilon T} \right] \frac{\theta_t}{n_i} \right| \right] \leq \sum_i \frac{1}{\sqrt{n_i}} + \text{ECE}(P) + T^{-\frac{1}{3}} \leq \text{ECE}(P) + 2T^{-\frac{1}{3}}. \quad \blacktriangleleft$$

## B.2 Proof of Theorem 27

We restate our lemmas for ECE and CDL separately here and prove them.

► **Theorem 33.** *For any post-processing algorithm  $f$ , there exists two sequences of predictions  $\mathbf{q}$  and  $\mathbf{q}'$  with states  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}'$ , respectively, both satisfying  $DTC(\mathbf{q}) = DTC(\mathbf{q}') = \epsilon$ , such that*

$$\max \{ \mathbf{E} [ECE(\mathbf{p}; \boldsymbol{\theta})], \mathbf{E} [ECE(\mathbf{p}'; \boldsymbol{\theta}')] \} \geq \frac{1}{8} \sqrt{\epsilon} + \frac{1}{2} \epsilon = \Theta(\sqrt{\epsilon}),$$

where we write  $\mathbf{p}, \mathbf{p}'$  as the output of the post-processing algorithm  $f$  on  $\mathbf{q}, \mathbf{q}'$ , respectively.

► **Lemma 34.** *Given predictor  $Q = (q_1, \dots, q_T)$ , and a post-processing algorithm  $f = (f_1, \dots, f_T)$ , suppose the empirical posterior for each prediction is  $\hat{Q} = (\hat{q}_1, \dots, \hat{q}_T)$ . There exists a sequence of states  $\boldsymbol{\theta}$  such that  $\boldsymbol{\theta}$  is compatible with the empirical posterior, i.e.*

$$\forall i \in [T], \hat{q}_i = \frac{\sum_{t \in [T]} \theta_t \mathbb{I}[q_t = q_i]}{\sum_{t \in [T]} \mathbb{I}[q_t = q_i]}.$$

Moreover, the expected ECE of the predictor  $f$  with states  $\boldsymbol{\theta}$  is lowerbounded

$$\mathbf{E}_{\mathbf{p} \sim f} [ECE(\mathbf{p}, \boldsymbol{\theta})] \geq \mathbf{E}_{\mathbf{p} \sim f} \left[ \frac{1}{T} \sum_{p \in \text{supp}(\mathbf{p})} \left| \sum_{t \in [T]} (p - \hat{q}_t) \cdot \mathbb{I}[p_t = p] \right| \right],$$

where  $\text{supp}$  is the support of the output of  $f$  in each round.

**Proof.** Define  $S_\theta = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \text{ is compatible with the empirical posterior}\}$ . Let  $\boldsymbol{\theta}$  be chosen uniformly at random from  $S_\theta$ , fix a sequence of predictions  $\mathbf{p}$ .

Given the distribution of  $\boldsymbol{\theta}$ ,  $\mathbf{E}_{\boldsymbol{\theta} \in S_\theta} [\sum_t \hat{p}_i \mathbb{I}[p_t = p_i]] = \sum_t \hat{q}_t \mathbb{I}[p_t = p_i]$  holds for any sequences of predictions  $\mathbf{p}$  and any  $i \in [T]$ . By Jensen's Inequality,

$$\begin{aligned} \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \left[ \left| \sum_t (p_i - \hat{p}_i) \mathbb{I}[p_t = p_i] \right| \right] &\geq \left| \sum_t (p_i \mathbb{I}[p_t = p_i] - \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} [\hat{p}_i \mathbb{I}[p_t = p_i]]) \right| \\ &= \left| \sum_t (p_i - \hat{q}_t) \mathbb{I}[p_t = p_i] \right|, \end{aligned}$$

apply this inequality to every prediction value  $p_t$ :

$$\begin{aligned} \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} [ECE(\mathbf{p})] &= \frac{1}{T} \sum_{p_i} \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \left[ \left| \sum_t (p_i - \hat{p}_i) \mathbb{I}[p_t = p_i] \right| \right] \\ &\geq \frac{1}{T} \sum_{p_i} \left| \sum_t (p_i - \hat{q}_t) \mathbb{I}[p_t = p_i] \right| \end{aligned}$$

Take expectation on the distribution of predictions,

$$\mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \mathbf{E}_{\mathbf{p} \sim P} [ECE(\mathbf{p})] \geq \mathbf{E}_{\mathbf{p} \sim f} \left[ \frac{1}{T} \sum_{p \in \text{supp}(\mathbf{p})} \left| \sum_{t \in [T]} (p - \hat{q}_t) \cdot \mathbb{I}[p_t = p] \right| \right].$$

Therefore, there must exist a sequence of states  $\boldsymbol{\theta}$  that

$$\mathbf{E}_{\mathbf{p} \sim P} [ECE(\mathbf{p})] \geq \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \mathbf{E}_{\mathbf{p} \sim P} [ECE(\mathbf{p})] \geq \mathbf{E}_{\mathbf{p} \sim f} \left[ \frac{1}{T} \sum_{p \in \text{supp}(\mathbf{p})} \left| \sum_{t \in [T]} (p - \hat{q}_t) \cdot \mathbb{I}[p_t = p] \right| \right].$$

◀

## 16:24 Smooth Calibration and Decision Making

**Proof of Theorem 33.** Assume there are  $2T$  rounds, define  $\mathbf{q}$  and  $\mathbf{q}'$  as following:

$$q_t = \begin{cases} \frac{1}{2} - \sqrt{\epsilon} & \text{if } t \leq T \\ \frac{1}{2} + \sqrt{\epsilon} & \text{else} \end{cases}$$

$$\sum_{t=1}^T \mathbb{I}[\theta_t = 1] = T \left( \frac{1}{2} - \frac{1}{2}\sqrt{\epsilon} + \epsilon \right), \quad \sum_{t=T+1}^{2T} \mathbb{I}[\theta_t = 1] = T \left( \frac{1}{2} + \frac{1}{2}\sqrt{\epsilon} - \epsilon \right).$$

For any  $t \in [2T]$ ,  $q'_t = \frac{1}{2} - \sqrt{\epsilon}$  and  $\sum_{t=1}^{2T} \mathbb{I}[\theta'_t = 1] = 2T \left( \frac{1}{2} - \sqrt{\epsilon} - \epsilon \right)$ . Define  $\hat{q}^0 = \frac{1}{2} - \sqrt{\epsilon} - \epsilon$ ,  $\hat{q}^1 = \frac{1}{2} - \frac{1}{2}\sqrt{\epsilon} + \epsilon$ ,  $\hat{q}^2 = \frac{1}{2} + \frac{1}{2}\sqrt{\epsilon} - \epsilon$ .

Fix a post-processing algorithm  $f$ . For any sequence of predictions  $\mathbf{p}$  generated by post-processing  $\mathbf{q}$ , denote the distribution of  $\mathbf{p}$  by  $\mathbf{f}(\mathbf{q})$ .

For any  $t' \in [2T]$  and any sequence of predictions  $\mathbf{p} \sim \mathbf{f}(\mathbf{q})$ , define

$$A(\mathbf{p})_{t'} = \frac{\hat{q}^1 \sum_{t=1}^T \mathbb{I}[p_t = p_{t'}] + \hat{q}^2 \sum_{t=T+1}^{2T} \mathbb{I}[p_t = p_{t'}]}{\sum_{t=1}^{2T} \mathbb{I}[p_t = p_{t'}]} \in [\hat{q}^1, \hat{q}^2].$$

According to Lemma 34, there exists a sequence of states  $\boldsymbol{\theta}$  that

$$\begin{aligned} \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} [\text{ECE}(\mathbf{p})] &\geq \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} \left[ \frac{1}{2T} \sum_{p \in \text{supp}(\mathbf{f}(\mathbf{q}))} \left| \sum_{t \in [2T]} (p - \hat{q}_t) \cdot \mathbb{I}[p_t = p] \right| \right] \\ &= \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} \left[ \frac{1}{2T} \sum_{t \in [2T]} |p_t - A(\mathbf{p})_t| \right]. \end{aligned} \quad (10)$$

According to Lemma 34,

$$\mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q}')} [\text{ECE}(\mathbf{p})] \geq \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q}')} \left[ \frac{1}{2T} \sum_{t \in [2T]} |p_t - \hat{p}^0| \right]. \quad (11)$$

For any  $\mathbf{p} \sim \mathbf{f}(\mathbf{q})$  and  $\mathbf{p}' \sim \mathbf{f}(\mathbf{q}')$ ,  $p_t = p'_t$  always holds for  $t \in [T]$ , since  $q_t = q'_t$  always holds for  $t \in [T]$ . Therefore, for any  $t \in [T]$ ,

$$|p_t - A(\mathbf{p})_t| + |p'_t - \hat{p}^0| \geq |A(\mathbf{p})_t - \hat{p}^0| \geq \frac{1}{2}\sqrt{\epsilon} + 2\epsilon.$$

Add up inequality (10) and inequality (11),

$$\begin{aligned} &\mathbf{E}_{\mathbf{p}' \sim \mathbf{f}(\mathbf{q}')} [\text{ECE}(\mathbf{p}')] + \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} [\text{ECE}(\mathbf{p})] \\ &\geq \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q}), \mathbf{p}' \sim \mathbf{f}(\mathbf{q}')} \left[ \frac{1}{2T} \sum_{t \in [2T]} (|p_t - A(\mathbf{p})_t| + |p'_t - \hat{p}^0|) \right] \\ &\geq \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q}), \mathbf{p}' \sim \mathbf{f}(\mathbf{q}')} \left[ \frac{1}{2T} \sum_{t \in [T]} (|p_t - A(\mathbf{p})_t| + |p'_t - \hat{p}^0|) \right] \\ &\geq \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q}), \mathbf{p}' \sim \mathbf{f}(\mathbf{q}')} \left[ \frac{1}{2T} \sum_{t \in [T]} |A(\mathbf{p})_t - \hat{q}^0| \right] \\ &= \frac{1}{4}\sqrt{\epsilon} + \epsilon. \end{aligned}$$

Therefore,

$$\begin{aligned} &\max \{ \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} [\text{ECE}(\mathbf{p})], \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q}')} [\text{ECE}(\mathbf{p})] \} \\ &\geq \frac{1}{2} \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} [\text{ECE}(\mathbf{p})] + \frac{1}{2} \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q}')} [\text{ECE}(\mathbf{p})] \geq \frac{1}{8}\sqrt{\epsilon} + \frac{1}{2}\epsilon. \end{aligned} \quad \blacktriangleleft$$

► **Theorem 35.** For any post-processing algorithm  $f$ , there exists two sequences of predictions  $\mathbf{q}$  and  $\mathbf{q}'$  with states  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}'$ , respectively, both satisfying  $DTC(\mathbf{q}) = DTC(\mathbf{q}') = \epsilon$ , such that

$$\max\{\mathbf{E}[ECE(\mathbf{p}; \boldsymbol{\theta})], \mathbf{E}[ECE(\mathbf{p}'; \boldsymbol{\theta}')]\} \geq \frac{1}{8}\sqrt{\epsilon} + \frac{1}{2}\epsilon = \Theta(\sqrt{\epsilon}),$$

where we write  $\mathbf{p}, \mathbf{p}'$  as the output of the post-processing algorithm  $f$  on  $\mathbf{q}, \mathbf{q}'$ , respectively.

Moreover, the same argument holds for CDL.

**Proof of Theorem 35.** Define two sets of sequences of states corresponding to predictor  $\mathbf{q}$  and  $\mathbf{q}'$  that every  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}'$  in these sets are compatible with empirical posterior:  $S_\theta = \{\boldsymbol{\theta} \mid \sum_{t \in [T]} \theta_t = T(\frac{1}{2} - \frac{1}{2}\sqrt{\epsilon} + \epsilon), \sum_{t=T+1}^{2T} \theta_t = T(\frac{1}{2} + \frac{1}{2}\sqrt{\epsilon} - \epsilon)\}$ ,  $S_{\theta'} = \{\boldsymbol{\theta} \mid \sum_{t \in [2T]} \theta_t = 2T(\frac{1}{2} - \sqrt{\epsilon} - \epsilon)\}$ . Denote the number of predicting prediction value  $p \in \text{supp}(\mathbf{p})$  by  $n_i = \sum_{t \in [2T]} \mathbb{I}[p_t = p_i]$ .

Fix a post-processing algorithm  $f$ . Define a proper scoring rule

$$S_\mu(p, \theta) = \begin{cases} \frac{1}{2} - \frac{1}{2} \cdot \frac{\theta - \mu}{\max\{\mu, 1 - \mu\}} & \text{if } p \leq \mu \\ \frac{1}{2} + \frac{1}{2} \cdot \frac{\theta - \mu}{\max\{\mu, 1 - \mu\}} & \text{else.} \end{cases}$$

According to the definition of CDL,

$$\mathbf{E}_{\mathbf{p} \sim f(\mathbf{q})}[\text{CDL}(\mathbf{p}, \boldsymbol{\theta})] \geq \frac{1}{2T} \mathbf{E}_{\mathbf{p} \sim f(\mathbf{q})} \left[ \sup_{\mu \in [0,1]} \sum_{t \in [2T]} (S_\mu(\hat{p}_t, \theta_t) - S_\mu(p_t, \theta_t)) \right]. \quad (12)$$

For any sequence of predictions  $\mathbf{p}$ , define  $N_{\mathbf{p}} = \sum_{t \in [T]} \mathbb{I}[p_t \geq \mu]$ ,  $M_{\mathbf{p}} = \sum_{t=T+1}^{2T} \mathbb{I}[p_t \geq \mu]$ .

$$\begin{aligned} & \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \left[ \sum_{t \in [2T]} (S_\mu(\hat{p}_t, \theta_t) - S_\mu(p_t, \theta_t)) \right] \\ & \geq \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \left[ \frac{1}{\max\{\mu, 1 - \mu\}} \sum_{p_i \in \text{supp}(\mathbf{p})} n_i \mathbb{I}[p_i \leq \mu] (\hat{p}_i - \mu) \right] \\ & = \frac{1}{\max\{\mu, 1 - \mu\}} \sum_{p_i \in \text{supp}(\mathbf{p}), p_i \leq \mu} \sum_{t \in [2T]} \mathbb{I}[p_t = p_i] (\hat{q}_t - \mu) \\ & \geq \frac{1}{\max\{\mu, 1 - \mu\}} \sum_{p_i \in \text{supp}(\mathbf{p}), p_i \leq \mu} \sum_{t \in [2T]} \mathbb{I}[p_t = p_i] \left( \frac{1}{2} - \frac{1}{2}\sqrt{\epsilon} + \epsilon - \mu \right) \\ & \geq \frac{1}{\max\{\mu, 1 - \mu\}} (2T - N_{\mathbf{p}} - M_{\mathbf{p}}) \left( \frac{1}{2} - \frac{1}{2}\sqrt{\epsilon} + \epsilon - \mu \right). \\ & \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \left[ \mathbf{E}_{\mathbf{p} \sim f(\mathbf{q})} \left[ \sup_{\mu \in [0,1]} \sum_{t \in [2T]} (S_\mu(\hat{p}_t, \theta_t) - S_\mu(p_t, \theta_t)) \right] \right] \\ & = \mathbf{E}_{\mathbf{p} \sim f(\mathbf{q})} \left[ \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \left[ \sup_{\mu \in [0,1]} \sum_{t \in [2T]} (S_\mu(\hat{p}_t, \theta_t) - S_\mu(p_t, \theta_t)) \right] \right] \\ & \geq \mathbf{E}_{\mathbf{p} \sim f(\mathbf{q})} \left[ \sup_{\mu \in [0,1]} \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \left[ \sum_{t \in [2T]} (S_\mu(\hat{p}_t, \theta_t) - S_\mu(p_t, \theta_t)) \right] \right] \\ & \geq \mathbf{E}_{\mathbf{p} \sim f(\mathbf{q})} \left[ \sup_{\mu \in [0,1]} \frac{1}{\max\{\mu, 1 - \mu\}} (2T - N_{\mathbf{p}} - M_{\mathbf{p}}) \left( \frac{1}{2} - \frac{1}{2}\sqrt{\epsilon} + \epsilon - \mu \right) \right]. \quad (13) \end{aligned}$$

## 16:26 Smooth Calibration and Decision Making

Combine inequality (12) and (13), there exists  $\boldsymbol{\theta} \in S_\theta$ , that

$$\begin{aligned}
& \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} [\text{CDL}(\mathbf{p}, \boldsymbol{\theta})] \\
& \geq \frac{1}{2T} \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} \left[ \sup_{\mu \in [0,1]} \frac{1}{\max\{\mu, 1-\mu\}} (2T - N_{\mathbf{p}} - M_{\mathbf{p}}) \left( \frac{1}{2} - \frac{1}{2}\sqrt{\epsilon} + \epsilon - \mu \right) \right]. \quad (14) \\
& \\
& \mathbf{E}_{\boldsymbol{\theta}' \in S_{\theta'}} \left[ \sum_{t \in [2T]} (S_\mu(\widehat{p}_t, \theta_t) - S_\mu(p_t, \theta_t)) \right] \\
& \geq \mathbf{E}_{\boldsymbol{\theta} \in S_\theta} \left[ \frac{1}{\max\{\mu, 1-\mu\}} \sum_{p_i \in \text{supp}(\mathbf{p}')} n_i \mathbb{I}[p_i \geq \mu] (\mu - \widehat{p}_i) \right] \\
& = \frac{1}{\max\{\mu, 1-\mu\}} \sum_{p_i \in \text{supp}(\mathbf{p}'), p_i \geq \mu} \sum_{t \in [2T]} \mathbb{I}[p_t = p_i] (\mu - \widehat{q}_t) \\
& = \frac{1}{\max\{\mu, 1-\mu\}} \sum_{p_i \in \text{supp}(\mathbf{p}'), p_i \geq \mu} \sum_{t \in [2T]} \mathbb{I}[p_t = p_i] \left( \mu - \frac{1}{2} + \sqrt{\epsilon} + \epsilon \right) \\
& = \frac{1}{\max\{\mu, 1-\mu\}} (N_{\mathbf{p}'} + M_{\mathbf{p}'}) \left( \mu - \frac{1}{2} + \sqrt{\epsilon} + \epsilon \right).
\end{aligned}$$

Similarly, there exists  $\boldsymbol{\theta}' \in S_{\theta'}$ , that

$$\begin{aligned}
& \mathbf{E}_{\mathbf{p}' \sim \mathbf{f}(\mathbf{q}')} [\text{CDL}(\mathbf{p}', \boldsymbol{\theta}')] \\
& \geq \frac{1}{2T} \mathbf{E}_{\mathbf{p}' \sim \mathbf{f}(\mathbf{q}')} \left[ \sup_{\mu \in [0,1]} \frac{1}{\max\{\mu, 1-\mu\}} (N_{\mathbf{p}'} + M_{\mathbf{p}'}) \left( \mu - \frac{1}{2} + \sqrt{\epsilon} + \epsilon \right) \right]. \quad (15)
\end{aligned}$$

For any  $\mathbf{p} \sim \mathbf{f}(\mathbf{q})$  and  $\mathbf{p}' \sim \mathbf{f}(\mathbf{q}')$ ,  $p_t = p'_t$  always holds for  $t \in [T]$ , since  $q_t = q'_t$  always holds for  $t \in [T]$ . So  $N_{\mathbf{p}} = N_{\mathbf{p}'}$  and  $M_{\mathbf{p}} = M_{\mathbf{p}'}$  always hold for  $t \in [T]$ . Combine inequality (14) and (15),

$$\begin{aligned}
& \max \{ \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} [\text{CDL}(\mathbf{p}, \boldsymbol{\theta})], \mathbf{E}_{\mathbf{p}' \sim \mathbf{f}(\mathbf{q}')} [\text{CDL}(\mathbf{p}', \boldsymbol{\theta}')] \} \\
& \geq \frac{1}{4T} \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q})} \left[ \sup_{\mu \in [0,1]} \frac{1}{\max\{\mu, 1-\mu\}} (2T - N_{\mathbf{p}} - M_{\mathbf{p}}) \left( \frac{1}{2} - \frac{1}{2}\sqrt{\epsilon} + \epsilon - \mu \right) \right] \\
& + \frac{1}{4T} \mathbf{E}_{\mathbf{p}' \sim \mathbf{f}(\mathbf{q}')} \left[ \sup_{\mu \in [0,1]} \frac{1}{\max\{\mu, 1-\mu\}} (N_{\mathbf{p}'} + M_{\mathbf{p}'}) \left( \mu - \frac{1}{2} + \sqrt{\epsilon} + \epsilon \right) \right] \quad (16)
\end{aligned}$$

$$\geq \frac{1}{4T} \mathbf{E}_{\mathbf{p} \sim \mathbf{f}(\mathbf{q}), \mathbf{p}' \sim \mathbf{f}(\mathbf{q}')} \left[ \frac{1}{\frac{1}{2} + \frac{3}{4}\sqrt{\epsilon}} (2T - M_{\mathbf{p}} + M_{\mathbf{p}'}) \left( \frac{1}{4}\sqrt{\epsilon} + \epsilon \right) \right] \quad (17)$$

$$\geq \frac{1}{8}\sqrt{\epsilon} + \frac{1}{2}\epsilon. \quad (18)$$

By taking  $\mu = \frac{1}{2} - \frac{3}{4}\sqrt{\epsilon}$  for both cases for  $\mathbf{p}$  and  $\mathbf{p}'$  and get (16)  $\geq$  (17). Since  $M_{\mathbf{p}}, M_{\mathbf{p}'} \in [0, T]$ ,  $M_{\mathbf{p}'} - M_{\mathbf{p}} \geq -T$ , so (17)  $\geq$  (18).  $\blacktriangleleft$