Testable and Actionable Calibration for Full Swap Regret

Konstantina Bairaktari
Department of Computer Science
Aarhus University

Lunjia Hu
Khoury College of Computer Sciences
Northeastern University

Huy L. Nguyen
Khoury College of Computer Sciences
Northeastern University

Jonathan Ullman
Khoury College of Computer Sciences
Northeastern University

Abstract

AI generated predictions increasingly inform decision making in critical tasks, and therefore must be trustworthy. One widely used measure of trustworthiness is calibration, which requires that the predictions match the true frequencies and can be treated like real probabilities of a given outcome. However, defining calibration is subtle, and designing good measures of calibration error has been an active topic of recent research. The first goal is to find calibration measures that are *actionable*, meaning they can inform decision makers about their utility loss when predictions are treated as true probabilities, which is known as swap regret. The second goal is to find calibration measures that are *testable*, meaning that calibration error can be measured from a small sample of predictions and outcomes. Although these are very basic requirements, there is no existing calibration measure that fully satisfies both properties, and all existing measures relax actionability by bounding a weaker notion of swap regret, or relax testability by having suboptimal estimation error. We introduce a new calibration measure, Soft-Binned Calibration Decision Loss (SCDL), which we prove is fully actionable without weakening either requirement, and testable with nearly optimal error rate. In addition, SCDL satisfies other desired properties such as continuity and consistency. We also provide a set of experiments confirming that the theoretical advantages of SCDL compared to other measures lead to better performance in practice.

1 Introduction

Deciding whether to carry an umbrella in the morning seems like a trivial problem, but it exemplifies the fundamental challenge of decision making under uncertainty, where we have to commit to an action without knowing the outcome in advance. To navigate this uncertainty, we turn to forecasts, a predicted probability of rain that compresses complex atmospheric dynamics into a single actionable number. But when can we trust this number? A forecast that reports a 20% chance of rain is meaningful if that number faithfully reflects the true probability of rain, but if it systematically overstates or understates the true probability, then decisions based on this forecast will be systematically wrong. The stakes grow considerably when we move from umbrellas to medical diagnoses where acting on an unreliable prediction can have severe consequences.

A decision maker can trust a predicted probability when it can be interpreted as a true probability and used to make the best possible decisions. Intuitively, when a forecast reports a 20% chance of rain, it should rain on 20% of the days that receive that forecast. This property, known as (perfect) calibration, has its roots in the classical forecasting literature [6]. Formally, given a predictive model

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

that outputs probabilities $p \in [0,1]$ for binary outcomes $y \in \{0,1\}$, perfect calibration requires that the joint distribution $\mathcal{D}$ of (p, y) satisfies

$$\mathbb{E}_{(p,y)\sim\mathcal{D}}[y \mid p] = p.$$

Perfect calibration is an ideal, it is never really achieved in practice, so the next question we need to ask is how far a predictor is from perfectly calibrated. Quantifying this calibration error turns out to be less straightforward than it may appear. A classic calibration measure is the *Expected Calibration Error (*ECE*)*, defined as the expected absolute difference between the prediction p and the true probability $\mathbb{E}[y|p]$, where the expectation is taken over the predictions made by the model. While the definition is very natural, ECE has the problem of being discontinuous and sensitive to small perturbations in the predictions, limiting its applicability for a robust and meaningful measure of calibration error. Many recent works aim to design better calibration measures that satisfy a variety of desired properties [1, 3, 2, 14, 22, 18, 36, 35, 19, 16, 17].

In this work, we build on the line of research that evaluates calibration through the lens of downstream decision making, where we evaluate calibration not based on the differences between the predicted and true probabilities, but by the impact these differences have on decision makers who use the predicted probabilities. In the literature, measures of calibration error that provide guarantees about the impact of miscalibration on the quality of downstream decisions are called *actionable* [36].

Measures of calibration error are a property of a predictor and a data distribution. But to be useful, the measure of calibration error must also be *testable*, meaning we can accurately estimate the calibration error using a small finite sample of predictions and outcomes. Without testability, we do not know which predictors actually perform well or how to improve our predictors to be more calibrated.

In this work, our goal is to design calibration measures that are simultaneously *fully actionable*, meaning they bound the impact of miscalibration on arbitrary downstream decision makers, and *nearly optimally testable*, meaning we can approximate the calibration error from samples with the best possible rate of convergence. Our main contribution is to introduce the first measure of calibration error that satisfies both properties. We emphasize that our focus is on the fundamental question of measuring the calibration error of a given predictor, rather than on learning a calibrated predictor.

1.1 An actionable calibration measure: Calibration Decision Loss

The quality of decisions can be evaluated through a utility function $U(a, y)$ that assigns a value to each combination of action a (take an umbrella or don't) and outcome y (rainy or not). A perfectly calibrated predictor allows the decision maker to act on the predicted probabilities as if they were true without suffering any *regret*, which is the feeling that they should have interpreted the predictions differently in hindsight, and swapped certain actions for others.

Example 1. *Consider a distribution $\mathcal{D}$ over (prediction, outcome), where the prediction is the probability of rain on some day and the outcome is 1 if it rains and 0 otherwise. The predictor says there is a 20% chance of rain with probability $\frac{1}{2}$ and says there is an 80% chance of rain with probability $\frac{1}{2}$. If the predictor is calibrated, then it rains on 20% of the days where the forecast is 20% and on 80% of the days with forecast 80%. Now suppose the decision maker carries an umbrella whenever the predicted probability exceeds 50%, and receives utility 1 for a correct match (carrying an umbrella on a rainy day or not carrying one on a dry day) and 0 otherwise. Under this strategy, the expected utility is $\frac{1}{2} \times 0.8 + \frac{1}{2} \times 0.8 = 0.80$. Since the decision maker has no information to go on other than the prediction, and chooses an action based solely on the predicted probability, a swap would mean choosing a different action for a given prediction value. In this case, no such swap improves the expected utility: carrying an umbrella when $p = 0.8$ and not carrying one when $p = 0.2$ is already optimal in hindsight.*

The connection between perfect calibration and decision making can be made precise: the joint distribution over predictions and outcomes is (perfectly) calibrated if and only if no decision maker incurs *swap regret* when relying solely on the predictions to take actions [8]. Let $\mathcal{D}$ be a joint distribution over predictions and outcomes $(p, y) \in [0,1] \times \{0,1\}$. For a decision task $Z = (A, U)$ with action space A and utility function U, the swap regret of a response function $r : [0,1] \to A$ is defined as

$$\mathsf{SR}_Z(r, \mathcal{D}) := \sup_{\sigma: A\to A} \mathbb{E}_{(p,y)\sim\mathcal{D}}[U(\sigma(r(p)), y) - U(r(p), y)]. \tag{1}$$

Upon receiving a prediction p, a decision maker who trusts it as the true probability will respond with the action $a \in A$ that maximizes the expected utility *assuming* that the outcome y is drawn from the Bernoulli distribution with mean p. Formally, the *best-response* function r_Z^* for a decision task $Z = (A, U)$ is defined as

$$r_Z^*(p) := \arg\max_{a \in A} \mathbb{E}_{y \sim \mathsf{Ber}(p)} U(a, y).$$

A distribution $\mathcal{D}$ of predictions and outcomes is calibrated if and only if for every decision task Z, the best-response function incurs no swap regret: $\mathsf{SR}_Z(r_Z^*, \mathcal{D}) = 0$. From the perspective of decision makers, this property defines what it means to be calibrated.

Thus, it is natural to measure the calibration error of a distribution $\mathcal{D}$ using the swap regret incurred for decision tasks. Hu and Wu [22] introduced a calibration error measure called the Calibration Decision Loss (CDL) as the swap regret $\mathsf{SR}_Z(r_Z^*, \mathcal{D})$ maximized over all decision tasks Z with a bounded utility function $U : A \times \{0, 1\} \to [0, 1]$. A distribution $\mathcal{D}$ is calibrated if and only if $\mathsf{CDL}(\mathcal{D}) = 0$. A low CDL provides the meaningful guarantee of low swap regret for any downstream task Z when the decision maker simply uses the best-response function r_Z^*. Because of this property, Rossellini et al. [36] describe CDL as an *actionable* measure of calibration.

1.2 The challenge of being both actionable and testable

Although CDL is actionable, it is fundamentally not *testable* [36], and cannot be accurately estimated from a finite dataset of predictions and outcomes. The intuition is that even a perfectly calibrated predictor can exhibit high CDL on a finite sample, so that simply measuring CDL on a finite sample cannot distinguish a perfectly calibrated predictor from a badly miscalibrated one. To be precise, when evaluating CDL on a sample S, we treat S as the uniform distribution over the data points in the sample. The following example shows that CDL on a finite sample can be bounded away from zero even when the underlying distribution is perfectly calibrated, for any sample size T.

Example 2. *Consider the decision task $Z = (A, U)$ where $A = [0, 1]$ and $U(a, y) = 1 - (a - y)^2$. The best-response function r_Z^* is the identity function $r_Z^*(p) = p$, so the swap regret for this decision task can be simplified as*

$$\mathsf{SR}_Z(r_Z^*, \mathcal{D}) = \sup_{\sigma:[0,1]\to[0,1]} \mathbb{E}[(p - y)^2 - (\sigma(p) - y)^2].$$

Let $\mathcal{D}$ be the calibrated distribution of (p, y) where p is drawn uniformly from $[1/3, 2/3]$ and y is drawn from $\mathsf{Ber}(p)$. On a finite sample S of T i.i.d. points $(p_1, y_1), \ldots, (p_T, y_T) \sim \mathcal{D}$, the predictions $p_1, \ldots, p_T$ are almost surely all distinct, and thus a swap function σ can map each $p_t \in [1/3, 2/3]$ to the corresponding $y_t \in \{0, 1\}$, incurring $\Omega(1)$ swap regret:

$$\mathsf{SR}_Z(r_Z^*, S) = \sup_{\sigma:[0,1]\to[0,1]} \frac{1}{T} \sum_{t=1}^{T} \left((p_t - y_t)^2 - (\sigma(p_t) - y_t)^2\right) = \frac{1}{T} \sum_{t=1}^{T} (p_t - y_t)^2 \geq 1/9.$$

Since $\mathsf{CDL}(S)$ is an upper bound on the swap regret $\mathsf{SR}_Z(r_Z^*, S)$, we also have $\mathsf{CDL}(S) \geq 1/9$.

Example 2 shows that a calibration error that is meaningful on the full distribution $\mathcal{D}$ may not be very meaningful on a finite sample S, no matter the sample size. In practice the predictor performance is almost always evaluated on finite data, and we need to be able to estimate the calibration error on new samples from the distribution based on the calibration error we observe on the sample.

We say a calibration measure CAL is *testable* if it incurs a small estimation error $|\mathsf{CAL}(\mathcal{D}) - \mathsf{CAL}(S)|$ between the full distribution $\mathcal{D}$ and a finite sample S.[1] Unfortunately, Example 2 shows that being testable is incompatible with providing the actionable no-swap-regret guarantee for the best-response function, even when we only require testability on perfectly calibrated distributions $\mathcal{D}$ in the limit of infinite sample size. Formally, Example 2 implies the following claim:

Claim 1.1. *No calibration measure CAL can satisfy the following two properties at the same time:*

1. ***Actionable for the best-response function.*** *For every distribution $\mathcal{D}$ of $(p, y) \in [0, 1] \times \{0, 1\}$ and every decision task $Z = (A, U)$ with bounded utility function $U : A \times \{0, 1\} \to [0, 1]$, we require $\mathsf{SR}_Z(r_Z^*, \mathcal{D}) \leq \mathsf{CAL}(\mathcal{D})$.*

[1] This requirement is stronger than the testability required by Rossellini et al. [36], since they allow an arbitrary estimator to be used on the sample S in order to estimate $\mathsf{CAL}(\mathcal{D})$, whereas we require the calibration measure itself to be an accurate plug-in estimator. For the calibration measures we discuss in this paper, the two notions give the same bounds.

2. ***Asymptotically testable for calibrated distributions.*** *Let $\mathcal{D}$ be a* calibrated *distribution of $(p, y) \in [0,1] \times \{0,1\}$. Let S_T be a sample of T i.i.d. points from $\mathcal{D}$. We require* $\lim_{T\to+\infty} \mathbb{E}[\mathsf{CAL}(S_T)] = 0$.

Thus, in order to achieve both actionability and testability, we will need to somehow relax one of the two properties. Specifically, we introduce and motivate a slight variant of actionability that allows us to circumvent this negative result, while preserving the spirit of actionability. In Section 1.3 we motivate our relaxation and compare it to other approaches to relaxing actionability.

Approaches to bypassing Claim 1.1 To overcome the impossibility from Claim 1.1, recent work shows that actionable and testable calibration measures can exist if we weaken actionability by considering only a restricted family of swap functions. Instead of taking the supremum over all swap functions $\sigma : A \to A$ as in (1), Rossellini, Soloff, Barber, Ren, and Willett [36] focus on swap functions that correspond to a monotone transformation of the predictions. They introduce the *cutoff calibration error* (Cutoff) and show that it is actionable for this restricted family of swap functions while being testable with a vanishing estimation error $O(T^{-1/2})$. Gopalan, Stavropoulos, Talwar, and Tankala [16] study calibration measures that are actionable w.r.t. a general family $\mathcal{K}$ of swap functions and show that the estimation error is characterized by the VC dimension of $\mathsf{thr}(\mathcal{K})$, the concept class consisting of thresholds applied to $\mathcal{K}$. If we were to include all swap functions in $\mathcal{K}$, then we would have the strict definition of actionability for the best-response function, but then the VC dimension of would be infinite and the measure would not be testable.

However, restricting the family of swap functions weakens the decision-theoretic meaning of calibration and introduces arbitrariness in the choice of the swap functions. Intuitively, considering only monotone swaps means we only measure miscalibration from a systemic upward or downward bias in the probabilities, and do not consider the possibility that the predictor may be miscalibrated in other ways such as systemic under or overconfidence.

1.3 Our approach: specifying the response function

Our goal is to design a testable and actionable calibration measure while preserving the strong decision-theoretic guarantee that we should be able to trust calibrated predictions and use them to make good decisions. The key idea, inspired by previous work such as [2, 20], is to change the response function in the definition of actionability to one that is more robust than the exact best-response function r^*_Z, and measure the swap regret with respect to that response function. For example, we use (a variant of) the response function that discretizes the predictions to some appropriate level of precision and then treats the discretized predictions as if they were real probabilities and applies best response. Intuitively, the goal of the strict definition of actionability is to ensure that the predictions can be used as if they were real probabilities to get strong decision-theoretic guarantees. By changing the response function to discretize the probabilities, we are instead saying that the high order bits of the predictions can be treated as if they were real probabilities and used to get strong decision-theoretic guarantees. Thus, this relaxation preserves the spirit of actionability by still giving a bound on the swap regret provided that the predictions are used in an appropriate way. We note that unlike restricting the class of swap functions, which makes difficult-to-verify assumptions about the predictor and its potential error, changing the response function only requires the user of the predictions to adjust their behavior in a way that is entirely under their own control.

We can see the value of this approach by revisiting Example 2. The swap regret in Example 2 is high because every prediction $p_1, \ldots, p_T$ is distinct, and can be mapped to a distinct action by the identity function $r^*_Z(p) = p$, and thus the swap regret is high because the decision-maker will only see one example of each predicted probability, and in hindsight might have chosen any sequence of actions. If we instead bucket the predictions and map predictions in the same bucket to the same action, the swap regret can be significantly reduced because the decision maker now sees many examples of the same bucket of predictions and will get a better sense of the true probability for the predictions in that bucket.

We view swap regret minimization as the *goal* of making calibrated predictions, so it is best to achieve the original goal rather than a weaker relaxation; we view the response function as the *method* used by decision makers to achieve the goal, so it can be flexible and need not be restricted to the exact best-response function.

Once we stop insisting on using the exact best-response function r_Z^*, the impossibility in Claim 1.1 no longer applies and we can hope for a testable and actionable calibration measure for full swap regret. We can formalize this objective in the following basic question.

Question 1. *Find a calibration measure* CAL_T *that satisfies the following properties:*[2]

1. ***Actionable for* some *response function.*** *For every decision task* $Z = (A, U)$ *with bounded utility function* $U : A \times \{0,1\} \to [0,1]$ *and every* $\alpha \geq 0$*, there exists a (possibly randomized)*[3] *response function* $r_{Z,\alpha}$ *such that for every distribution* $\mathcal{D}$ *on* $[0,1] \times \{0,1\}$ *satisfying* $\mathsf{CAL}_T(\mathcal{D}) \leq \alpha$*, the decisions made by applying* $r_{Z,\alpha}$ *to the prediction have low swap regret. Formally, it holds that* $\mathsf{SR}_Z(r_{Z,\alpha}, \mathcal{D}) \leq \alpha$*. Importantly, the response function cannot depend on the distribution* $\mathcal{D}$*, which the decision maker does not know.*

2. $\varepsilon(T)$***-testable for calibrated distributions.*** *Let* $\mathcal{D}$ *be a* calibrated *distribution on* $[0,1] \times \{0,1\}$*. Let* S_T *be a sample of* T *i.i.d. points from* $\mathcal{D}$*. It holds that* $\mathbb{E}[\mathsf{CAL}_T(S_T)] \leq \varepsilon(T)$*. We want* $\varepsilon(T)$ *as small as possible.*

Intuitively, the strictest definition of actionability requires the response function to be exact best-response, which interprets the prediction p as an exact probability and then chooses a response by calculating the utility-maximizing action, and requires that decisions made in this way have low swap regret. However, this strict definition leads to the impossibility result of Claim 1.1. In our relaxation, we get to specify some alternative way of using the predicted probabilities to make decisions and provide bounds on the swap regret of those decisions. As discussed above, our alternative response function will roughly be to discretize the probabilities to a suitable precision and then apply best response, which we believe is a natural way of interpreting predicted probabilities.

Prior work on answers to Question 1. Many existing calibration measures satisfy the two properties in Question 1 with vanishing estimation error $\varepsilon(T) \to 0$ as $T \to +\infty$. The best error rate we can get using measures proposed in prior work is $\varepsilon(T) = O(T^{-1/3})$, achieved by a binned variant of ECE. Specifically, we choose $m \approx T^{1/3}$ and define a discretization function $b_m : [0,1] \to \{0, 1/m, 2/m, \ldots, 1\}$ that rounds predictions $p \in [0,1]$ down to the nearest multiple of $1/m$. Given a distribution $\mathcal{D}$ of prediction-outcome pairs (p, y), we let $\mathcal{D}_m$ denote the distribution of the discretized predictions and outcomes $(b_m(p), y)$. We define *binned ECE* (bECE_m) by applying ECE to the discretized distribution $\mathcal{D}_m$ and multiplying the result by 2. That is, $\mathsf{bECE}_m(\mathcal{D}) := 2\,\mathsf{ECE}(\mathcal{D}_m)$.

This definition of bECE_m is actionable when decision makers use a response function r_Z that is slightly different from the best-response function r_Z^*. Specifically, when a decision maker sees a prediction p, they should apply the best-response function r_Z^* *to the discretized prediction* $b_m(p)$. That is, we use $r_Z := r_Z^* \circ b_m$ as the response function for decision task Z. It is known that $2\,\mathsf{ECE}$ is an upper bound on CDL, so $\mathsf{bECE}_m(\mathcal{D}) = 2\,\mathsf{ECE}(\mathcal{D}_m)$ is a valid upper bound on the swap regret $\mathsf{SR}_Z(r_Z^*, \mathcal{D}_m)$, which is equal to $\mathsf{SR}_Z(r_Z, \mathcal{D})$ because the discretization b_m in $r_Z = r_Z^* \circ b_m$ turns $\mathcal{D}$ into $\mathcal{D}_m$. This shows that bECE satisfies the actionability requirement of Question 1.

The testability $\varepsilon(T) = O(T^{-1/3})$ of bECE_m can be proved by showing that the expected value of $\mathsf{bECE}_m(S)$ for a sample S of T i.i.d. points from any calibrated distribution $\mathcal{D}$ is $O(T^{-1/3})$. Roughly, this can be obtained by summing up the discretization error $O(1/m) = O(T^{-1/3})$ and a sampling error which is of order $O(\sqrt{m/T}) = O(T^{-1/3})$. Choosing $m \approx T^{1/3}$ balances the two terms and achieves the optimal testability for binned ECE. The discretization b_m can be replaced by adding random noise to the predictions, where the absolute value of the noise is around $1/m \approx T^{-1/3}$. This leads to a calibration measure called *smooth ECE* introduced by [2] that is also actionable and achieves the same $O(T^{-1/3})$-testability.

In contrast, the cutoff calibration error [36] has stronger testability $O(T^{-1/2})$, but it is not fully actionable because it only bounds swap regret for monotone swaps. For general swap functions, Roth and Shi [37] show an actionability guarantee of the cutoff calibration error that degrades with the action space size $|A|$. That is, the swap regret in a downstream decision task with action space A and

[2] To make the question easier, we allow the calibration measure to depend on the sample size T. This dependency is often not needed. The calibration measure SCDL we introduce will not depend on T.

[3] We formally define the swap regret of a randomized response function in Definition D.4. The definition is in-expectation, but our calibration measure SCDL will satisfy a stronger high-probability guarantee as well.

Table 1: SCDL achieves full actionability and near-optimal testability. Blue text denotes the most desirable guarantee, while black text denotes a suboptimal guarantee.

Measure	Actionability	Testability
CDL	Full	$\Theta(1)$
smCE	None	$\Theta(T^{-1/2})$
$\text{smCE}^{1/2}$	Full	$\Theta(T^{-1/4})$
Cutoff	Monotone	$\Theta(T^{-1/2})$
Cutoff^{α}	Full only for $\alpha \leq 2/3$	$\Theta(T^{-\alpha/2}) = \Omega(T^{-1/3})$
Binned/Smooth ECE	Full	$\Theta(T^{-1/3})$
SCDL (Our Work)	Full	$O(T^{-1/2} \log T)$

a bounded utility function $U : A \times \{0, 1\} \to [0, 1]$ is at most $O(|A|)$ times the cutoff calibration error. Due to the dependency on $|A|$, this is weaker than the actionability guarantee in Question 1, where we require a uniform upper bound for all utility-bounded decision tasks independent of the number of actions. Our goal is to achieve the same testability as cutoff calibration error while satisfying the stronger actionability requirement of Question 1, with no dependency on A and no restriction on the swap family.

The smooth calibration error smCE [26, 3] also achieves the desired $O(T^{-1/2})$-testability. However, it is actionable only *after we take its square root* [20, 17]. Taking the square root compromises its testability from $O(T^{-1/2})$ to $O(T^{-1/4})$, which becomes even worse than the $O(T^{-1/3})$-testability guarantee of binned ECE. The square root is indeed necessary to make the smooth calibration error actionable, even when general response functions are allowed. This is shown by an analysis in [20]. We present a simplified version of this analysis in Appendix I. Similarly, in Theorem J.1 we show that if one wishes to make the cutoff calibration error fully actionable by taking its α-th power, it is necessary to take $\alpha \leq 2/3$ even when arbitrary response functions are allowed, compromising its testability from $O(T^{-1/2})$ to $O(T^{-1/3})$. For a summary of calibration measures satisfying Question 1, see Table 1.

1.4 Our results: a calibration measure with nearly optimal actionability and testability

Our first contribution is an improved solution to Question 1 achieving $O(T^{-1/2} \log T)$-testability. We also prove a lower bound of $\Omega(T^{-1/2})$ showing that our solution is tight up to a logarithmic factor (Theorem H.1).

We prove our upper bound by designing a new calibration measure which we term *Soft-Binned Calibration Decision Loss (*SCDL*)*. In addition to solving Question 1, we show that SCDL satisfies the following desired properties:

1. SCDL is **testable** not just for calibrated distributions, but for every distribution $\mathcal{D}$. For a sample S of T i.i.d. points drawn from $\mathcal{D}$,
$$\mathbb{E}|\mathsf{SCDL}(S) - \mathsf{SCDL}(\mathcal{D})| \leq O(T^{-1/2} \log^{3/2} T). \tag{2}$$
Moreover, the testability of SCDL holds not just in expectation, but also with high probability. We formally state and prove the testability of SCDL in Theorem G.1. When $\mathsf{SCDL}(\mathcal{D}) = 0$ we prove a slightly stronger bound that $\mathbb{E}[\mathsf{SCDL}(S)] = O(T^{-1/2} \log T)$.
2. SCDL is **actionable** for the full swap regret with a simple response function $r_{Z,\alpha}$. Similarly to our earlier discussion about binned ECE, we define $r_{Z,\alpha}(p)$ by applying the best-response function r^*_Z to a discretized version p' of the given prediction p. The discretization is very basic and does not depend on the decision task Z. We discuss the actionability of SCDL in more detail in Section 2.
3. SCDL is **sound and complete**. That is, a distribution $\mathcal{D}$ satisfies $\mathsf{SCDL}(\mathcal{D}) = 0$ if and only if it is perfectly calibrated.
4. SCDL is a **consistent** calibration measure in the framework of [3]. That is, SCDL is polynomially, in fact quadratically, related to the *lower distance to calibration* DistCal

(Theorem E.1):

$$\Omega(\mathsf{DistCal}(\mathcal{D})^2) \le \mathsf{SCDL}(\mathcal{D}) \le O(\sqrt{\mathsf{DistCal}(\mathcal{D})}). \tag{3}$$

The lower distance to calibration DistCal (Definition B.4) is introduced by [3] as a sound and complete calibration measure. Thus the quadratic relationship above implies the soundness and completeness of SCDL. We also show that both the upper and lower bounds in (3) are tight up to constant factors (Remark E.1).

5. SCDL is **continuous** as a function of the predictions. That is, small changes to the predictions can only cause small changes in SCDL. In Theorem F.1, we prove that SCDL is $\frac{1}{2}$-Hölder continuous.

We complement our theoretical results with experiments following the setup of [36]. On the testability side, SCDL exhibits the smallest estimation variance across most settings, suggesting it can be estimated more reliably from finite samples than the other measures. On the actionability side, we evaluate the correlation between each calibration measure and swap regret across a range of prediction distributions. For a predictor that corresponds to a monotone optimal swap function, SCDL, ECE and Cutoff correlate well with swap regret. However, when the optimal swap function is not monotone, Cutoff loses its correlation with swap regret while SCDL continues to track it, illustrating the advantage of not restricting the swap function class.

Related work. Although we have discussed the key related work in the introduction, we defer more exhaustive coverage of related work to Appendix A.

2 Actionability and Testability of SCDL

Definition of SCDL. We define SCDL as a soft-binned variant of CDL. Similarly to how binned ECE is actionable and testable, the actionability of SCDL is inherited from the actionability of CDL, whereas the testability of SCDL comes from binning, which is a natural idea that has been used widely in practice. Our main technical contribution lies in proving the concrete, near-optimal testability of SCDL using a self-balancing binning granularity, which we discuss below.

We start by defining the quantity SCDL_m, for a positive integer m. First we discretize predictions into a finite grid $G_m := \{0, 1/m, 2/m, \dots, 1\}$, where we view each grid point as a bin. Specifically, given a prediction p, we randomly round it to the closest multiple of $1/m$ either above or below, which we call p', and we choose the probability of rounding up or down so that $\mathbb{E}p' = p$. For every distribution $\mathcal{D}$ of $(p, y) \in [0, 1] \times \{0, 1\}$, we use $\mathcal{D}_m$ to denote the distribution of $(p', y) \in G_m \times \{0, 1\}$ where we use rounded predictions.

Given this discretization, the natural choice is to define $\mathsf{SCDL}_m(\mathcal{D}) = \mathsf{CDL}(\mathcal{D}_m)$ to be the calibration decision loss of the discretized predictions. While this is almost what we do, we make two technical modifications. The first modification is to replace $\mathsf{CDL}(\mathcal{D}_m)$ with an explicit formula that gives a factor-2 approximation to CDL and is easier to analyze (Lemma C.4). The second modification is to remove small bin-wise prediction bias up to $1/m$. We make the second modification because the discretization can cause even a perfectly calibrated distribution to have nonzero $\mathsf{CDL}(\mathcal{D}_m)$, and we want to subtract this rounding error off so that $\mathsf{SCDL}_m(\mathcal{D}) = 0$ when the predictions are perfectly calibrated. This modification makes $\mathsf{SCDL}_m(\mathcal{D})$ non-decreasing as $m = 2^k$ grows as a power of two, which will help us choose the right discretization granularity m to establish testability.

For any choice of m, the quantity $2(\mathsf{SCDL}_m(\mathcal{D}) + 1/m)$ satisfies the actionability requirement of Question 1. To see why this is the case, first we have $\mathsf{SCDL}_m(\mathcal{D}) \approx \mathsf{CDL}(\mathcal{D}_m)$, and $\mathsf{CDL}(\mathcal{D}_m)$ is a valid upper bound on the swap regret $\mathsf{SR}_Z(r_Z^*, \mathcal{D}_m)$ of the best-response function r_Z^* applied to the discretized distribution $\mathcal{D}_m$. Next, the response function we define will take a prediction p, discretize it in the same way to obtain p', and then apply the best-response function to p'. Therefore, the swap regret of this new reponse function on the original distribution $\mathcal{D}$ is the same as the swap regret of the best-response function on $\mathcal{D}_m$. The factor 2 is due to the approximation formula for CDL and the extra $1/m$ term is to correct for the rounding error that we removed from the definition of SCDL_m.

To complete the definition of SCDL, it remains to choose the value of m appropriately. While the choice of m does not affect the actionability of $2(\mathsf{SCDL}_m(\mathcal{D}) + 1/m)$, it makes a crucial difference in terms of testability. Indeed, CDL can be viewed as a special case of SCDL_m with $m = +\infty$, but CDL

is not testable. It turns out that to obtain our $\varepsilon(T) = O(T^{-1/2} \log T)$-testability guarantee, it suffices to choose $m \approx 1/\varepsilon(T)$. This choice, however, has the disadvantage of depending on the sample size T. Also, as we will discuss shortly, it only provides testability for calibrated distributions $\mathcal{D}$ and does not satisfy the testability in (2) for general distributions. Instead, we choose m in a self-balancing way so that $\mathsf{SCDL}_m(\mathcal{D})$ matches the rounding error $1/m$. Concretely, we choose $m = 2^k$ as the power of 2 that minimizes $\max\{\mathsf{SCDL}_m(\mathcal{D}), 1/m\}$ and define $\mathsf{SCDL}(\mathcal{D})$ as this minimum value. We show that $\mathsf{SCDL}_{2^k}(\mathcal{D})$ is an increasing function of k whereas $1/m$ is a decreasing function. Thus, by minimizing the maximum of the two quantities, our choice of m ensures a balance between them, and the final error $\mathsf{SCDL}(\mathcal{D})$ is guaranteed to lie in the interval $[1/m, 2/m)$. This idea of choosing m to balance $1/m$ and the binned calibration error is inspired by the *interval calibration error* from [3] and it turns out to be essential for establishing our testability guarantees.

Formally, we define SCDL as follows.

Definition 2.1 (Soft-Binned CDL $\mathsf{SCDL}(\mathcal{D})$)**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0, 1] \times \{0, 1\}$. For every positive integer m, we define $\mathcal{D}_m$ as the distribution of (p', y), where p' is obtained by randomly rounding p up or down to the closest multiple of $1/m$ while ensuring $\mathbb{E}[p'|p, y] = p$. For every $i = 0, \ldots, m$, define the weight π_i and the true conditional probability q_i as follows:*

$$\pi_i := \Pr_{(p',y)\sim\mathcal{D}_m}[p' = i/m], \quad q_i := \mathbb{E}_{(p',y)\sim\mathcal{D}_m}[y|p' = i/m].$$

For every $i = 0, \ldots, m$, define the m-soft-binned CDL $\mathsf{SCDL}_m(\mathcal{D})$ as the following quantity that approximates $\mathsf{CDL}(\mathcal{D}_m)$ (see Lemma B.2), where x_+ denotes $\max\{x, 0\}$:

$$\mathsf{SCDL}_m(\mathcal{D}) := \max_{i=0,\ldots,m} \left(\sum\nolimits_{j=0,\ldots,i} \pi_j (q_j - (i+1)/m)_+ + \sum\nolimits_{j=i+1,\ldots,m} \pi_j (i/m - q_j)_+ \right).$$

Finally, we define the soft-binned CDL of $\mathcal{D}$ by choosing m as a power of two that balances $1/m$ and $\mathsf{SCDL}_m(\mathcal{D})$. That is, $\mathsf{SCDL}(\mathcal{D}) := \inf_{m\in\{2^1,2^2,\ldots\}} \max\{\mathsf{SCDL}_m(\mathcal{D}), 1/m\}$.

Actionability of SCDL. As discussed, the actionability of SCDL follows naturally from its definition. Suppose we are given a predictor where the joint distribution $\mathcal{D}$ of (p, y) satisfies $\mathsf{SCDL}(\mathcal{D}) \le \alpha$. Here is how we use it to achieve low swap regret on a decision task $Z = (A, U)$ with a bounded utility function $U : A \times \{0, 1\} \to [0, 1]$. If $\alpha = 0$, then $\mathcal{D}$ is guaranteed to be perfectly calibrated and we simply use the best-response function $r_{Z,0} := r^*_Z$ to achieve zero swap regret $\mathsf{SR}_Z(r^*_Z, \mathcal{D}) = 0$. If $\alpha > 0$, we find $m \in \{2^1, 2^2, \ldots, \}$ such that $\alpha \in [1/m, 2/m)$ and randomly round the prediction p up or down to the closest multiple p' of $1/m$ while ensuring $\mathbb{E}[p'] = p$. We take the best action $a = r^*_Z(p')$ based on the rounded prediction p'. That is, $r_{Z,\alpha}(p) := r^*_Z(p')$. We prove the following swap regret guarantee in Theorem D.2:

$$\mathsf{SR}_Z(r_{Z,\alpha}, \mathcal{D}) \le 2\,\mathsf{SCDL}(\mathcal{D}) + 2/m \le 4\alpha. \tag{4}$$

This implies that SCDL satisfies the full actionability required in Question 1. Note that we incur an additional factor of 4, so we need to set $\mathsf{CAL} = 4\,\mathsf{SCDL}$ in Question 1 to satisfy the actionability requirement as stated. We choose not to include this factor of 4 in the definition of SCDL for a clearer presentation.

The response function $r_{Z,\alpha}$ that achieves the swap regret guarantee (4) is randomized due to randomness in the rounding procedure. We formally define the swap regret of a randomized response function $r : [0, 1] \to \Delta(A)$ in Definition D.4, where the regret against each swap function $\sigma : A \to A$ is calculated in expectation over actions sampled from $r(p)$. In Theorem D.4, we show that the actionablity guarantee of SCDL holds not just in expectation, but also with high probability.

We remark that we can replace the randomized rounding in the definition of SCDL with a deterministic rounding procedure, where every p is rounded to the largest multiple p' of $1/m$ that does not exceed p. This "hard-binned" variant preserves all the testability and actionability properties of SCDL following the same analysis in our work. However, this variant does not enjoy the continuity of SCDL. It is an interesting question for future research to understand the power of other response functions that do not require a discretization of the prediction space, such as quantal response functions and other smoothed best-response functions.

Testability of SCDL. Proving the testability of SCDL is the most technically challenging part of our work. To bound the estimation error $|\mathsf{SCDL}(S) - \mathsf{SCDL}(\mathcal{D})|$ between the full distribution $\mathcal{D}$ and a random sample S, the first step is to bound the estimation error for the conditional expectation

$\mathbb{E}[y|p' = i/m]$ for each bin i/m, where p' is the rounded version of p. Roughly speaking, a standard concentration argument shows that on average, each bin incurs an estimation error in the order of $\sqrt{m/T}$. Using this analysis, to get our final testability guarantee around $T^{-1/2}$, we are only allowed to have roughly $m = O(1)$ bins, in which case the rounding error $1/m$ will dominate, resulting in a large testability error.

To improve this basic analysis, we use a key result from Hu and Wu [22] who showed that for CDL, the errors from all bins do not contribute equally to the final estimation error. When $\mathcal{D}$ is calibrated, they show that, roughly, only a $\sqrt{m/T}$ fraction of the bin-wise estimation errors contribute to the final error. This analysis shows that the estimation error reduces to roughly m/T. When we choose $m \approx T^{1/2}$, this m/T estimation error and the rounding error $1/m$ are both of order $T^{-1/2}$.

However, this argument is limited to calibrated distributions $\mathcal{D}$, and substantial new ideas are needed to prove the testability result for a general distribution $\mathcal{D}$. We show that as $\mathsf{SCDL}(\mathcal{D})$ grows, the estimation error deteriorates smoothly by roughly an additive $O(\sqrt{m \cdot \mathsf{SCDL}(\mathcal{D})/T})$ term. Importantly, since $\mathsf{SCDL}(\mathcal{D}) \leq 1$, this bound is better than the naive bound of $O(\sqrt{m/T})$. However, for a highly miscalibrated distribution $\mathcal{D}$ with $\mathsf{SCDL}(\mathcal{D}) = \Omega(1)$, the two bounds are of the same order and we can no longer choose $m \approx T^{1/2}$ to get the desired testability bound. Instead, we would like to choose m as a function of $\mathsf{SCDL}(\mathcal{D})$, so that when $\mathsf{SCDL}(\mathcal{D})$ is large, we can compensate for higher estimation error by taking fewer bins. This is accomplished by our self-balancing choice of m, which ensures that $m \cdot \mathsf{SCDL}(\mathcal{D}) = O(1)$ for *all* distributions $\mathcal{D}$. Plugging this into the error bound gives the desired $\approx \sqrt{1/T}$ error. We summarize this result in the following theorem:

Theorem 2.1 (SCDL is testable, see Appendix G for proof of a stronger high-probability result)**.** *Let S be a sample of $T \geq 2$ data points $(p_1, y_1), \ldots, (p_T, y_T)$ drawn i.i.d. from a distribution $\mathcal{D}$ on $[0,1] \times \{0,1\}$. Then, in expectation over S, we have the bound $\mathbb{E}|\mathsf{SCDL}(S) - \mathsf{SCDL}(\mathcal{D})| \leq O(T^{-1/2} \log^{3/2} T)$. Moreover, when $\mathsf{SCDL}(\mathcal{D}) = 0$, we have $\mathbb{E}[\mathsf{SCDL}(S)] = O(T^{-1/2} \log T)$.*

3 Experiments

We follow the experimental setup of [36] and extend it with additional experiments.[4] For a given $\alpha \in [0,1]$, each sample (X_i, Y_i) is drawn as follows: the feature X_i is sampled independently from the uniform distribution on $[0,1]$, and the label Y_i is drawn from a Bernoulli distribution with mean $\mathbb{E}[Y|X = x] = \alpha(1-2x)^2 + (1-\alpha)x$. We fit a predictor f using logistic regression on $n = 500$ samples from the distribution. To break monotonicity, we also consider the predictor that flips the labels, i.e. $1 - f$.

Testability. In Table 2, we compute the value and the standard deviation of the estimated calibration errors from $N = 1000$ samples. To estimate ECE, we use bECE for 10 bins of equal length.

Table 2: Calibration error estimates (mean $\pm$ std) for predictor f across values of α.

Measure	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.8$	$\alpha = 1$
smCE	0.021 ± 0.014	0.028 ± 0.013	0.027 ± 0.016	0.025 ± 0.016
Cutoff	0.030 ± 0.012	0.068 ± 0.016	0.110 ± 0.016	0.136 ± 0.015
bECE	0.043 ± 0.011	0.117 ± 0.015	0.140 ± 0.054	0.064 ± 0.065
SCDL	0.016 ± 0.003	0.036 ± 0.006	0.080 ± 0.014	0.076 ± 0.034

Across most values of α in Table 2, SCDL has the smallest standard deviation among the four error measures, suggesting it can be estimated more reliably from finite samples. The exception is $\alpha = 1$, where smCE and Cutoff have lower variance, though SCDL still compares favorably to bECE.

Actionability. We evaluate actionability using the threshold utility from [36], with each probability treated as a distinct action: the action space is $A = [0,1]$, the utility function is $U(a,y) = 1 - 0.35(1-y)\mathbb{I}[a \geq 0.35] - 0.65y\mathbb{I}[a < 0.35]$, and we choose the best-response function $r^*_Z(p) = p$. Although actions on the same side of the threshold have identical utilities, a swap function $\sigma : A \to A$

[4] Code for our experiments is available at [https://github.com/konstantinabrk/scdl].

may map them to different sides of the threshold. The swap-regret benchmark therefore chooses the better of the two utility types separately for each prediction. For 200 uniformly random choices of α, we compute each calibration measure and the swap regret of the best-response action using the same setup as above. To keep the comparison fair across measures, we compute swap regret using the original, non-discretized best-response function r^*_Z for every measure, including SCDL. This is a more demanding test for SCDL. Our actionability guarantee holds for the discretized response function $r_{Z,\alpha}$, so this experiment checks whether SCDL remains informative even when the decision maker does not discretize the predictions. We show scatter plots of the results for f in Figure 3 (Appendix K) and for $1 - f$ in Figure 1, where darker shades of blue correspond to higher α. Following [36], we quantify the alignment between each calibration error measure and the swap regret using the Spearman correlation, where a higher correlation means that the calibration error more faithfully informs the decision maker about their swap regret. While actionability is a one-sided guarantee, upper bounding the swap regret simultaneously over all bounded decision tasks, an individual decision maker would also like the calibration error to indicate the swap regret on their own task. While our theoretical results establish the upper bound, this experiment complements them by measuring, on a single fixed decision task as in [36], how well each calibration error tracks the swap regret.

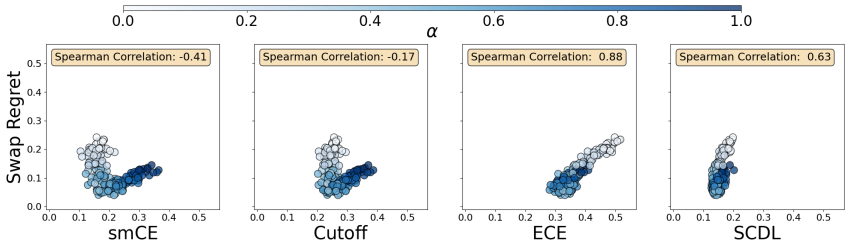


Figure 1: Plots of calibration error measures versus swap regret for predictor $1 - f$. Spearman correlations are reported in each subplot.

For predictor f, all measures except smCE are highly correlated with swap regret (see Figure 3). However, the picture changes for $1 - f$. Cutoff is no longer correlated with swap regret, it just upper bounds it, with several (prediction, outcome) distributions sharing similar swap regret but different Cutoff values. ECE remains highly correlated and SCDL shows moderate correlation, indicating that it tracks swap regret adequately well. This can be explained by the fact that the optimal swap function for $1 - f$ is not monotone, falling outside the swap function class that Cutoff restricts to when bounding the swap regret. These results support our approach of relaxing actionability by discretizing the predictions rather than by restricting the swap functions.

Acknowledgments

The authors thank Yifan Wu for insightful discussions that initiated this work as well as helpful conversations about [20]. We also thank Parikshit Gopalan for conversations about [16, 17].

KB was supported by the European Union (ERC, TUCLA, 101125203). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. KB was also supported by NSF awards CNS-2232692 and CCF-2311649, while at Northeastern University.

HN was supported by NSF award CCF-2311649.

JU was supported by NSF awards CNS-2232692 and CNS-2247484.

References

[1] I. Arrieta-Ibarra, P. Gujral, J. Tannen, M. Tygert, and C. Xu. Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*, 23(351):1–54, 2022. URL http://jmlr.org/papers/v23/22-0658.html. 2

[2] J. Blasiok and P. Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XwiA1nDahv. 2, 4, 5

[3] J. Błasiok, P. Gopalan, L. Hu, and P. Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 1727–1740, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399135. doi: 10.1145/3564246.3585182. URL https://doi.org/10.1145/3564246.3585182. 2, 6, 7, 8, 13, 14, 15

[4] J. Blasiok, P. Gopalan, L. Hu, and P. Nakkiran. When does optimizing a proper loss yield calibration? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 72071–72095. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e4165c96702bac5f4962b70f3cf2f136-Paper-Conference.pdf. 13

[5] Y. Dagan, C. Daskalakis, M. Fishelson, N. Golowich, R. Kleinberg, and P. Okoroafor. Breaking the Tˆ(2/3) barrier for sequential calibration. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, page 2007–2018, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718178. URL https://doi.org/10.1145/3717823.3718178. 14

[6] A. P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982. doi: 10.1080/01621459.1982.10477856. 1

[7] M. Fishelson, N. Golowich, M. Mohri, and J. Schneider. High-dimensional calibration from swap regret. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=UVDihUz0iT. 14

[8] D. P. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(589):40–55, 1997. 2, 14

[9] D. P. Foster and R. V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 06 1998. ISSN 0006-3444. doi: 10.1093/biomet/85.2.379. URL https://doi.org/10.1093/biomet/85.2.379. 14

[10] S. Garg, C. Jung, O. Reingold, and A. Roth. Oracle efficient online multicalibration and omniprediction. In D. P. Woodruff, editor, *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, pages 2725–2792. SIAM, 2024. doi: 10.1137/1.9781611977912.98. URL https://doi.org/10.1137/1.9781611977912.98. 13

[11] P. Gopalan, A. T. Kalai, O. Reingold, V. Sharan, and U. Wieder. Omnipredictors. In M. 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. ISBN 978-3-95977-217-4. doi: 10.4230/LIPIcs.ITCS.2022.79. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2022.79. 13

[12] P. Gopalan, L. Hu, M. P. Kim, O. Reingold, and U. Wieder. Loss Minimization Through the Lens Of Outcome Indistinguishability. In Y. Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 60:1–60:20, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-263-1. doi: 10.4230/LIPIcs.ITCS.2023.60. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2023.60. 13

[13] P. Gopalan, M. Kim, and O. Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. In A. Oh, T. Naumann, 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. 13

[14] P. Gopalan, L. Hu, and G. N. Rothblum. On computationally efficient multi-class calibration. In S. Agrawal and A. Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1983–2026. PMLR,

30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/gopalan24a.html. 2

[15] P. Gopalan, P. Okoroafor, P. Raghavendra, A. Sherry, and M. Singhal. Omnipredictors for regression and the approximate rank of convex functions. In S. Agrawal and A. Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2027–2070. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/gopalan24b.html. 13

[16] P. Gopalan, K. Stavropoulos, K. Talwar, and P. Tankala. Efficient calibration for decision making. *arXiv preprint arXiv:2511.13699*, 2025. 2, 4, 10

[17] P. Gopalan, K. Stavropoulos, K. Talwar, and P. Tankala. The importance of being smoothly calibrated. *arXiv preprint arXiv:2603.16015*, 2026. 2, 6, 10, 15

[18] N. Haghtalab, M. Qiao, K. Yang, and E. Zhao. Truthfulness of calibration measures. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 117237–117290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3722. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d4cbcae8cfc8aa3ae897a1296e4e0cac-Paper-Conference.pdf. 2, 14

[19] J. Hartline, L. Hu, and Y. Wu. A perfectly truthful calibration measure. *arXiv preprint arXiv:2508.13100*, 2025. 2, 14

[20] J. Hartline, Y. Wu, and Y. Yang. Smooth Calibration and Decision Making. In M. Bun, editor, *6th Symposium on Foundations of Responsible Computing (FORC 2025)*, volume 329 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 16:1–16:26, Dagstuhl, Germany, 2025. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-367-6. doi: 10.4230/LIPIcs.FORC.2025.16. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.FORC.2025.16. 4, 6, 10, 28

[21] U. Hebert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In J. Dy and A. 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 Jul 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html. 13

[22] L. Hu and Y. Wu. Predict to minimize swap regret for all payoff-bounded tasks. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 244–263, 2024. doi: 10.1109/FOCS61266.2024.00024. 2, 3, 9, 14, 20

[23] L. Hu, I. R. Livni Navon, O. Reingold, and C. Yang. Omnipredictors for constrained optimization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. 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 Jul 2023. URL https://proceedings.mlr.press/v202/hu23b.html. 13

[24] L. Hu, K. Tian, and C. Yang. Omnipredicting single-index models with multi-index models. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, page 1762–1773, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718223. URL https://doi.org/10.1145/3717823.3718223. 13

[25] L. Hu, K. Tian, and C. Yang. Simultaneous blackwell approachability and applications to multiclass omniprediction. *arXiv preprint arXiv:2602.17577*, 2026. 13

[26] S. M. Kakade and D. P. Foster. Deterministic calibration and Nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2007.04.017. URL https://www.sciencedirect.com/science/article/pii/S0022000007000633. Learning Theory 2004. 6, 14

[27] M. P. Kim and J. C. Perdomo. Making Decisions Under Outcome Performativity. In Y. Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:15, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-263-1. doi: 10.4230/LIPIcs.ITCS.2023.79. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2023.79. 13

[28] M. P. Kim, C. Kern, S. Goldwasser, F. Kreuter, and O. Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022. doi: 10.1073/pnas.2108097119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2108097119. 13

[29] B. Kleinberg, R. P. Leme, J. Schneider, and Y. Teng. U-calibration: Forecasting for an unknown agent. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5143–5145. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/kleinberg23a.html. 14

[30] J. Lu, A. Roth, and M. Shi. Sample efficient omniprediction and downstream swap regret for non-linear losses. In N. Haghtalab and A. Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3829–3878. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/lu25b.html. 13

[31] G. Noarov, R. Ramalingam, A. Roth, and S. Xie. High-dimensional unbiased prediction for sequential decision making. In *OPT 2023: Optimization for Machine Learning*, 2023. URL https://openreview.net/forum?id=P4j4l45NUq. 13

[32] P. Okoroafor, R. Kleinberg, and M. P. Kim. Near-optimal algorithms for omniprediction. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1595–1609, 2025. doi: 10.1109/FOCS63196.2025.00084. 13, 14

[33] B. Peng. High dimensional online calibration in polynomial time. *arXiv preprint arXiv:2504.09096*, 2025. 14

[34] M. Qiao and G. Valiant. Stronger calibration lower bounds via sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 456–466, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3451050. URL https://doi.org/10.1145/3406325.3451050. 14

[35] M. Qiao and E. Zhao. Truthfulness of decision-theoretic calibration measures. In N. Haghtalab and A. Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4686–4739. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/qiao25a.html. 2, 14

[36] R. Rossellini, J. A. Soloff, R. F. Barber, Z. Ren, and R. Willett. Can a calibration metric be both testable and actionable? In N. Haghtalab and A. Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4937–4972. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/rossellini25a.html. 2, 3, 4, 5, 7, 9, 10, 14, 15

[37] A. Roth and M. Shi. Forecasting for swap regret for all downstream agents. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 466–488, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707049. doi: 10.1145/3670865.3673622. URL https://doi.org/10.1145/3670865.3673622. 5, 14

A Related work

Recent research on calibration measures is largely motivated by Błasiok, Gopalan, Hu, and Nakkiran [3], who studied the design of calibration measures in a systematic way. They introduce the distance to calibration as a central notion and identifies other calibration measures as *consistent* if they are polynomially related to the distance to calibration. Our calibration measure SCDL is quadratically related to the distance to calibration (3), so it is consistent according to the definition in [3]. Consistent calibration measures have been used to explain practical phenomena in deep learning (see e.g. [4]).

Closely related to our goal of swap regret minimization for downstream decision makers, *omniprediction* is the task of making predictions that allow a large family of downstream decision makers to achieve better or similar utility compared to a family of benchmark decision functions [11]. Although omniprediction does not require a swap regret guarantee, notions of calibration have been extensively used for constructing omnipredictors [12, 23, 27, 31, 13, 15, 10, 24, 32, 30, 25]. Besides omniprediction, calibration and its variants have played important roles in the study of algorithmic fairness [21] and distributional robustness [28].

Hu and Wu [22] introduced the decision-theoretic calibration measure CDL motivated by the classic question of making calibrated sequential predictions [9, 34, 5, 33, 7] and the task of simultaneous online regret minimization for all downstream decision tasks [29, 37]. Since [22], many recent works aim to find decision-theoretically meaningful calibration measures with additional properties. Besides the work on testability which we have discussed, Qiao and Zhao [35] introduce a decision-theoretic calibration measure that is also approximately *truthful*, where a truthful calibration measure incentivizes predictors to report true probabilities [18, 19]. The step calibration error used in [35] can be viewed as equivalent to the cutoff calibration error from [36], although the two works are concurrent and independent. The same calibration measure was also independently found to be useful for achieving omniprediction [32].

B Notation and Background

For every $x \in \mathbb{R}$, we use x_+ to denote $\max\{x, 0\}$. We use $\mathbb{I}[E]$ to denote the 0-1 indicator for a statement E, where $\mathbb{I}[E] = 1$ if E is true, and $\mathbb{I}[E] = 0$ if E is false. For every $p \in [0, 1]$, we use $\mathsf{Ber}(p)$ to denote the Bernoulli distribution with mean p. We use $\Delta(A)$ to denote the set of all probability distributions on an action set A.

Below we present the formal definitions of existing calibration measures that are most related to our work. We also include key properties of these measures that will be useful for proving properties of SCDL.

Definition B.1 (Expected Calibration Error [8])**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0, 1] \times \{0, 1\}$. The Expected Calibration Error of $\mathcal{D}$ is defined as $\mathsf{ECE}(\mathcal{D}) = \mathbb{E}_{\mathcal{D}}|q - p|$, where q is the conditional expectation $\mathbb{E}_{\mathcal{D}}[y \mid p]$, which is a function of p.*

Definition B.2 (Calibration Decision Loss [22])**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0, 1] \times \{0, 1\}$. The Calibration Decision Loss of $\mathcal{D}$ is defined as*

$$\mathsf{CDL}(\mathcal{D}) := \sup_Z \mathsf{SR}_Z(r^*_Z, \mathcal{D}),$$

where the supremum is over all decision tasks $Z = (A, U)$ with an arbitrary action space A and a bounded utility function $U : A \times \{0, 1\} \to [0, 1]$.

The following relationship between CDL and ECE is established in [22]:

Lemma B.1 ([22])**.** *For every distribution $\mathcal{D}$ on $[0, 1] \times \{0, 1\}$, it holds that*

$$\mathsf{ECE}(\mathcal{D})^2 \le \mathsf{CDL}(\mathcal{D}) \le 2\,\mathsf{ECE}(\mathcal{D}).$$

The following lemma gives an explicit formula for computing CDL up to a factor of 2:

Lemma B.2 ([22])**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0, 1] \times \{0, 1\}$. Let $q = \mathbb{E}[y|p] \in [0, 1]$ be a function of p representing the conditional expectation of y given p. We have*

$$\mathsf{CDL}(\mathcal{D}) \le 2 \sup_{\mu \in [0,1]} \mathbb{E}[(q - \mu)_+ \mathbb{I}[p \le \mu] + (\mu - q)_+ \mathbb{I}[p > \mu]].$$

We also have the reverse inequality up to a factor of 2:

$$\mathsf{CDL}(\mathcal{D}) \ge \sup_{\mu \in [0,1]} \mathbb{E}[(q - \mu)_+ \mathbb{I}[p \le \mu] + (\mu - q)_+ \mathbb{I}[p > \mu]].$$

Definition B.3 (Smooth Calibration Error [26, 3])**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0, 1] \times \{0, 1\}$. The smooth calibration error $\mathsf{smCE}(\mathcal{D})$ is defined as*

$$\mathsf{smCE}(\mathcal{D}) := \sup_{\text{1-Lipschitz } w\,:\,[0,\,1] \to [-1,\,1]} \mathbb{E}_{(p,y)\sim\mathcal{D}}[(y - p)w(p)],$$

where the supremum is over all 1-Lipschitz functions $w : [0, 1] \to [-1, 1]$.

Definition B.4 (Lower Distance to Calibration [3])**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0, 1] \times \{0, 1\}$. The lower distance to calibration $\mathsf{DistCal}(\mathcal{D})$ is defined as*

$$\mathsf{DistCal}(\mathcal{D}) := \inf_{\mathcal{J}} \mathbb{E}_{(p,p',y)\sim\mathcal{J}}[|p - p'|],$$

where the infimum is over all joint distributions $\mathcal{J}$ of $(p, p', y) \in [0, 1] \times [0, 1] \times \{0, 1\}$ such that the distribution of (p, y) is $\mathcal{D}$ and the distribution of (p', y) is calibrated.

The following theorems shows that smCE and DistCal are constant-factor approximations of each other:

Theorem B.3 ([3, 17])**.** *For every distribution $\mathcal{D}$ on $[0,1] \times \{0,1\}$,*

$$\frac{1}{2}\,\mathsf{DistCal}(\mathcal{D}) \le \mathsf{smCE}(\mathcal{D}) \le 2\,\mathsf{DistCal}(\mathcal{D}).$$

Definition B.5 (Cutoff calibration error [36])**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0,1] \times \{0,1\}$. The cutoff calibration error $\mathsf{Cutoff}(\mathcal{D})$ is defined as*

$$\mathsf{Cutoff}(\mathcal{D}) := \sup_{0 \le a \le b \le 1} |\mathbb{E}_{(p,y)\sim\mathcal{D}}[(y-p)\mathbb{I}[a \le p \le b]]|.$$

C Basic Properties of SCDL

In this section we prove some basic properties of SCDL. These properties will be useful for us to show its actionability, testability, consistency, and continuity in later sections.

The following lemma gives an equivalent definition of the bin-wise weight π_i and true conditional probability q_i used in our definition of SCDL (Definition 2.1).

Lemma C.1. *Let m be a positive integer. For every $i = 0, 1, \ldots, m$ and $p \in [0,1]$, define $w_i(p) := (1 - |mp - i|)_+ \in [0,1]$. For every distribution $\mathcal{D}$ of $(p, y) \in [0,1] \times \{0,1\}$ and every $i = 0, \ldots, m$, the π_i and q_i from Definition 2.1 satisfies*

$$\pi_i := \mathbb{E}_{(p,y)\sim\mathcal{D}}[w_i(p)], \quad q_i := \frac{1}{\pi_i}\mathbb{E}_{(p,y)\sim\mathcal{D}}[w_i(p)y].$$

Proof. Let p' be the rounded prediction in Definition 2.1. We have $\Pr[p' = i/m | p] = w_i(p)$. Plugging it into the definition of π_i and q_i completes the proof. □

It will be convenient to define the following intermediate notion for a single bin i:

Definition C.1 ($\mathsf{SCFDL}_{m,i}(\mathcal{D})$)**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0,1] \times \{0,1\}$. Let m be a positive integer. For every $i = 0, 1, \ldots, m$, define π_i and q_i as in Definition 2.1. For every $i = 0, \ldots, m$, define the* soft-binned calibration fixed decision loss *as follows:*

$$\mathsf{SCFDL}_{m,i}(\mathcal{D}) := \sum_{j=0}^{i} \pi_j (q_j - (i+1)/m)_+ + \sum_{j=i+1}^{m} \pi_j (i/m - q_j)_+.$$

Combining Definition C.1 and Definition 2.1, we have

$$\mathsf{SCDL}_m(\mathcal{D}) = \max_{i=0,\ldots,m} \mathsf{SCFDL}_{m,i}(\mathcal{D}).$$

The following lemma shows that $\mathsf{SCDL}_m(\mathcal{D})$ is a non-decreasing function of m when m is a power of 2.

Lemma C.2. *For every positive integer m and every distribution $\mathcal{D}$ on $[0,1] \times \{0,1\}$, $\mathsf{SCDL}_{2m}(\mathcal{D}) \ge \mathsf{SCDL}_m(\mathcal{D})$.*

Proof. For the discretization in $2m$ bins, for every $i = 0, 1, \ldots, 2m$, define $\tilde{w}_i(p) := (1 - |2mp - i|)_+$, $\tilde{\pi}_i := \mathbb{E}_{(p,y)\sim\mathcal{D}}[\tilde{w}_i(p)]$ and $\tilde{q}_i := \frac{1}{\tilde{\pi}_i}\mathbb{E}_{(p,y)\sim\mathcal{D}}[\tilde{w}_i(p)y]$. Then, we see that for any $p \in [0,1]$

$$w_i(p) = \begin{cases} \tilde{w}_0(p) + \frac{1}{2}\tilde{w}_1(p) & i = 0, \\ \tilde{w}_{2i}(p) + \frac{1}{2}\tilde{w}_{2i-1}(p) + \frac{1}{2}\tilde{w}_{2i+1}(p) & 1 \le i \le m-1, \\ \tilde{w}_{2m}(p) + \frac{1}{2}\tilde{w}_{2m-1}(p) & i = m. \end{cases}$$

Therefore, by taking the expectation of $w_i(p)$ we have that

$$\pi_i = \begin{cases} \tilde{\pi}_0 + \frac{1}{2}\tilde{\pi}_1 & i = 0, \\ \tilde{\pi}_{2i} + \frac{1}{2}\tilde{\pi}_{2i-1} + \frac{1}{2}\tilde{\pi}_{2i+1} & 1 \le i \le m-1, \\ \tilde{\pi}_{2m} + \frac{1}{2}\tilde{\pi}_{2m-1} & i = m. \end{cases}$$

Similarly, taking the expectation of $w_i(p)y$, we obtain

$$q_i = \begin{cases} \frac{1}{\pi_i}(\tilde{\pi}_0\tilde{q}_0 + \frac{1}{2}\tilde{\pi}_1\tilde{q}_1) & i = 0, \\ \frac{1}{\pi_i}(\tilde{\pi}_{2i}\tilde{q}_{2i} + \frac{1}{2}\tilde{\pi}_{2i-1}\tilde{q}_{2i-1} + \frac{1}{2}\tilde{\pi}_{2i+1}\tilde{q}_{2i+1}) & 1 \le i \le m-1, \\ \frac{1}{\pi_i}(\tilde{\pi}_{2m}\tilde{q}_{2m} + \frac{1}{2}\tilde{\pi}_{2m-1}\tilde{q}_{2m-1}) & i = m. \end{cases}$$

For a fixed $i \in \{0, \dots, m\}$ functions $(x - (i+1)/m)_+$ and $(i/m - x)_+$ are convex in x. Therefore, by Jensen's inequality for a fixed $i \in \{0, \dots, m\}$ we have that

$$\begin{aligned}
&\mathsf{SCFDL}_{m,i}(\mathcal{D}) \\
&= \sum_{j=0}^{i} \pi_j \left(q_j - \frac{i+1}{m}\right)_+ + \sum_{j=i+1}^{m} \pi_j \left(\frac{i}{m} - q_j\right)_+ \\
&\le \tilde{\pi}_0 \left(\tilde{q}_0 - \frac{i+1}{m}\right)_+ + \frac{1}{2}\tilde{\pi}_1 \left(\tilde{q}_1 - \frac{i+1}{m}\right)_+ \\
&\quad + \sum_{j=1}^{i} \left(\frac{1}{2}\tilde{\pi}_{2j-1} \left(\tilde{q}_{2j-1} - \frac{i+1}{m}\right)_+ + \tilde{\pi}_{2j} \left(\tilde{q}_{2j} - \frac{i+1}{m}\right)_+ + \frac{1}{2}\tilde{\pi}_{2j+1} \left(\tilde{q}_{2j+1} - \frac{i+1}{m}\right)_+ \right) \\
&\quad + \sum_{j=i+1}^{m-1} \left(\frac{1}{2}\tilde{\pi}_{2j-1} \left(\frac{i}{m} - \tilde{q}_{2j-1}\right)_+ + \tilde{\pi}_{2j} \left(\frac{i}{m} - \tilde{q}_{2j}\right)_+ + \frac{1}{2}\tilde{\pi}_{2j+1} \left(\frac{i}{m} - \tilde{q}_{2j+1}\right)_+ \right) \\
&\quad + \frac{1}{2}\tilde{\pi}_{2m-1} \left(\frac{i}{m} - \tilde{q}_{2m-1}\right)_+ + \tilde{\pi}_{2m} \left(\frac{i}{m} - \tilde{q}_{2m}\right)_+ \\
&= \sum_{j=0}^{2i} \tilde{\pi}_j \left(\tilde{q}_j - \frac{i+1}{m}\right)_+ + \frac{1}{2}\tilde{\pi}_{2i+1} \left(\tilde{q}_{2i+1} - \frac{i+1}{m}\right)_+ \\
&\quad + \frac{1}{2}\tilde{\pi}_{2i+1} \left(\frac{i}{m} - \tilde{q}_{2i+1}\right)_+ + \sum_{j=2i+2}^{2m} \tilde{\pi}_j \left(\frac{i}{m} - \tilde{q}_j\right)_+ \\
&= \sum_{j=0}^{2i} \tilde{\pi}_j \left(\tilde{q}_j - \frac{2i+2}{2m}\right)_+ + \frac{1}{2}\tilde{\pi}_{2i+1} \left(\tilde{q}_{2i+1} - \frac{2i+2}{2m}\right)_+ \\
&\quad + \frac{1}{2}\tilde{\pi}_{2i+1} \left(\frac{2i}{2m} - \tilde{q}_{2i+1}\right)_+ + \sum_{j=2i+2}^{2m} \tilde{\pi}_j \left(\frac{2i}{2m} - \tilde{q}_j\right)_+ \\
&\le \frac{1}{2}\mathsf{SCFDL}_{2m,2i}(\mathcal{D}) + \frac{1}{2}\mathsf{SCFDL}_{2m,2i+1}(\mathcal{D}) \\
&\le \max\{\mathsf{SCFDL}_{2m,2i}(\mathcal{D}), \mathsf{SCFDL}_{2m,2i+1}(\mathcal{D})\}.
\end{aligned}$$

Hence, we conclude that

$$\mathsf{SCDL}_m(\mathcal{D}) = \max_{i=0,\dots,m} \mathsf{SCFDL}_{m,i}(\mathcal{D}) \le \max_{i=0,\dots,2m} \mathsf{SCFDL}_{2m,i}(\mathcal{D}) = \mathsf{SCDL}_{2m}(\mathcal{D}).$$

□

In our definition of SCDL, we choose m by minimizing $\max\{\mathsf{SCDL}_m(\mathcal{D}), 1/m\}$. We give a characterization of the minimizing m in Definition C.2 and Lemma C.3.

Definition C.2 (Optimal Number of Bins)**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0,1] \times \{0,1\}$. We define $m^*(\mathcal{D})$ as the smallest number m among $2^1, 2^2, \dots$, such that $\mathsf{SCDL}_{2m}(\mathcal{D}) \ge 1/m$. When such m does not exist, we define $m^*(\mathcal{D}) = +\infty$.*

Lemma C.3. *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0,1] \times \{0,1\}$. If $m^* := m^*(\mathcal{D}) = +\infty$, then $\mathsf{SCDL}(\mathcal{D}) = 0$. If $m^* < +\infty$, we have*

1. *$\mathsf{SCDL}(\mathcal{D}) = \max\{\mathsf{SCDL}_{m^*}(\mathcal{D}), 1/m^*\} \in [1/m^*, 2/m^*)$;*

2. $\mathsf{SCDL}_{2m^*}(\mathcal{D}) \geq \mathsf{SCDL}(\mathcal{D})$.

Proof. If $m^* = +\infty$, by the monotonicity in Lemma C.2, we have that for all m in $2^1, 2^2, \ldots$ $\mathsf{SCDL}_m(\mathcal{D}) \leq \mathsf{SCDL}_{2m}(\mathcal{D}) < 1/m$. Hence, $\mathsf{SCDL}(\mathcal{D}) = \inf_{m=2^1,2^2,\ldots} \max\{\mathsf{SCDL}_m(\mathcal{D}), 1/m\} = \inf_{m=2^1,2^2,\ldots} 1/m = 0$.

Now, we analyze the case for $m^* < +\infty$. First, we show that $\max\{\mathsf{SCDL}_{m^*}(\mathcal{D}), 1/m^*\} \in [1/m^*, 2/m^*)$. If $m^* \geq 4$, since m^* is the smallest number among $2^1, 2^2, \ldots$ such that $\mathsf{SCDL}_{2m^*}(\mathcal{D}) \geq 1/m^*$, we have that for $m^*/2$

$$\mathsf{SCDL}_{m^*}(\mathcal{D}) < \frac{2}{m^*}.$$

Therefore, if $\mathsf{SCDL}_{m^*}(\mathcal{D}) \leq 1/m^*$, then $\max\{\mathsf{SCDL}_{m^*}(\mathcal{D}), 1/m^*\} = 1/m^*$. Otherwise, $\max\{\mathsf{SCDL}_{m^*}(\mathcal{D}), 1/m^*\} = \mathsf{SCDL}_{m^*}(\mathcal{D}) < \frac{2}{m^*}$. For $m^* = 2$, we notice that $\mathsf{SCDL}_2(\mathcal{D}) \leq 1/2$. Thus, $\max\{\mathsf{SCDL}_2(\mathcal{D}), 1/2\} = 1/2 \in [1/m^*, 2/m^*)$.

Next, we show that $\mathsf{SCDL}(\mathcal{D}) = \max\{\mathsf{SCDL}_{m^*}(\mathcal{D}), 1/m^*\}$. For m that is dyadic and $m > m^*$, by Lemma C.2, we have that $\mathsf{SCDL}_m(\mathcal{D}) \geq \mathsf{SCDL}_{2m^*}(\mathcal{D}) \geq 1/m^*$ and $\mathsf{SCDL}_m(\mathcal{D}) \geq \mathsf{SCDL}_{m^*}(\mathcal{D})$. Therefore, $\max\{\mathsf{SCDL}_m(\mathcal{D}), 1/m\} \geq \max\{\mathsf{SCDL}_{m^*}(\mathcal{D}), 1/m^*\}$. For m that is dyadic and $m < m^*$, it holds that $m \leq m^*/2$. Thus, $\max\{\mathsf{SCDL}_m(\mathcal{D}), 1/m\} \geq 1/m \geq 2/m^* > \max\{\mathsf{SCDL}_{m^*}(\mathcal{D}), 1/m^*\}$.

By Lemma C.2, $\mathsf{SCDL}_{2m^*}(\mathcal{D}) \geq \mathsf{SCDL}_{m^*}(\mathcal{D})$ and by the definition of m^*, we have that $\mathsf{SCDL}_{2m^*}(\mathcal{D}) \geq 1/m^*$. Thus, $\mathsf{SCDL}_{2m^*}(\mathcal{D}) \geq \mathsf{SCDL}(\mathcal{D})$. □

The following lemma shows a basic relationship between SCDL and CDL.

Lemma C.4. *Let $\mathcal{D}$ be a distribution on $[0,1] \times \{0,1\}$. For every positive integer m, it holds that $\mathsf{SCDL}_m(\mathcal{D}) \leq \mathsf{CDL}(\mathcal{D})$. Consequently, $\mathsf{SCDL}(\mathcal{D}) \leq \mathsf{CDL}(\mathcal{D})$.*

Proof. Let π_i and q_i be defined as in Definition 2.1 for $i = 0, \ldots, m$. To show $\mathsf{SCDL}_m(\mathcal{D}) \leq \mathsf{CDL}(\mathcal{D})$, it suffices to show that for every $i = 0, \ldots, m$,

$$\sum_{j=0}^{i} \pi_j (q_j - (i+1)/m)_+ + \sum_{j=i+1}^{m} \pi_j (i/m - q_j)_+ \leq \mathsf{CDL}(\mathcal{D}). \tag{5}$$

Consider the joint distribution of $(p, y, q, p') \in [0,1] \times \{0,1\} \times [0,1] \times \{0, 1/m, \ldots, 1\}$, where $(p, y) \sim \mathcal{D}$, p' is the rounded version of p as in Definition 2.1, and $q := \mathbb{E}[y|p]$. We have

$$q_j = \mathbb{E}[y|p' = j/m] = \mathbb{E}\Big[\mathbb{E}[y|p' = j/m, p]\Big|p' = j/m\Big] = \mathbb{E}[q|p' = j/m],$$

where the first equation is the law of total probability, and the second equation holds because y is independent of p' conditioned on p.

Since x_+ is a convex function of x, by Jensen's inequality,

$$(q_j - (i+1)/m)_+ \leq \mathbb{E}[(q - (i+1)/m)_+ | p' = j/m],$$
$$(i/m - q_j)_+ \leq \mathbb{E}[(i/m - q)_+ | p' = j/m].$$

Therefore,

$$\begin{aligned} &\sum_{j=0}^{i} \pi_j (q_j - (i+1)/m)_+ + \sum_{j=i+1}^{m} \pi_j (i/m - q_j)_+ \\ &\leq \mathbb{E}[(q - (i+1)/m)_+ \mathbb{I}[p' \leq i/m]] + \mathbb{E}[(i/m - q)_+ \mathbb{I}[p' \geq (i+1)/m]]. \end{aligned} \tag{6}$$

For every $p \in [0,1]$, define

$$\beta(p) := \Pr[p' \leq i/m | p] = \begin{cases} 1, & \text{if } p \leq (i-1)/m, \\ i + 1 - pm, & \text{if } p \in (i/m, (i+1)/m), \\ 0, & \text{if } p \geq (i+1)/m. \end{cases}$$

We have $\Pr[p' \geq (i+1)/m | p] = 1 - \beta(p)$. By (6),

$$\begin{aligned}
&\sum_{j=0}^{i} \pi_j (q_j - (i+1)/m)_+ + \sum_{j=i+1}^{m} \pi_j (i/m - q_j)_+ \\
&\leq \mathbb{E}[(q - (i+1)/m)_+ \beta(p)] + \mathbb{E}[(i/m - q)_+ (1 - \beta(p))]. \qquad (7)
\end{aligned}$$

Let μ be drawn from the uniform distribution on $[i/m, (i+1)/m]$, independent of (p, q, p'). For every $p \in [0,1]$, we have

$$\Pr_{\mu}[p \leq \mu] = \beta(p), \quad \Pr_{\mu}[p > \mu] = 1 - \beta(p). \qquad (8)$$

By Lemma B.2,

$$\begin{aligned}
&\mathsf{CDL}(\mathcal{D}) \\
&\geq \mathbb{E}_{\mu} \mathbb{E}[(q-\mu)_+ \mathbb{I}[p \leq \mu] + (\mu - q)_+ \mathbb{I}[p > \mu]] \\
&\geq \mathbb{E}_{\mu} \mathbb{E}[(q - (i+1)/m)_+ \mathbb{I}[p \leq \mu] + (i/m - q)_+ \mathbb{I}[p > \mu]] \\
&= \mathbb{E}[(q - (i+1)/m)_+ \beta(p) + (i/m - q)_+ (1 - \beta(p))], \qquad (9)
\end{aligned}$$

where we used (8) to obtain the last equality. Combining (7) and (9) proves (5).

Now, we show that $\mathsf{SCDL}(\mathcal{D}) \leq \mathsf{CDL}(\mathcal{D})$. By Lemma C.3, we have that if $m^* = +\infty$, then $\mathsf{SCDL}(\mathcal{D}) = 0 \leq \mathsf{CDL}(\mathcal{D})$. Otherwise, we have that

$$\mathsf{SCDL}(\mathcal{D}) \leq \mathsf{SCDL}_{2m^*}(\mathcal{D}) \leq \mathsf{CDL}(\mathcal{D}). \qquad \square$$

D Actionability of SCDL

We formally state and the prove the actionability of SCDL in this section.

For every positive integer m, let G_m denote the grid $\{0, 1/m, \ldots, 1\}$. When defining SCDL in Definition 2.1, we rounded each prediction $p \in [0,1]$ to a grid point $p' \in G_m$. We formally define this rounding function as follows:

Definition D.1 (Randomized rounding function τ_m)**.** *For every prediction $p \in [0,1]$, we let $\tau_m(p)$ be the distribution of the rounded prediction $p' \in G_m$ obtained by randomly rounding p up or down to the closest multiple of $1/m$ while ensuring $\mathbb{E}[p'|p] = p$.*

Definition D.2 (Rounded distribution $\mathcal{D}_m$)**.** *Given (p, y) drawn from a distribution $\mathcal{D}$ on $[0,1] \times \{0,1\}$, we draw $p' \sim \tau_m(p)$ and use $\mathcal{D}_m$ to denote the distribution of (p', y).*

The following lemma follows directly by combining Definition 2.1 and Definition D.2.

Lemma D.1. *Let $\mathcal{D}$ be a distribution on $[0,1] \times \{0,1\}$ and let $\mathcal{D}_m$ be defined as in Definition D.2. We have*

$$\mathsf{SCDL}_m(\mathcal{D}_m) = \mathsf{SCDL}_m(\mathcal{D}).$$

Definition D.3 (Randomized response function $r \circ \tau_m$)**.** *For every response function $r : [0,1] \to A$, we use $r \circ \tau_m : [0,1] \to \Delta(A)$ to denote the randomized response function where $r \circ \tau_m(p)$ is the distribution of $r(p')$ for $p' \sim \tau_m(p)$.*

Definition D.4 (Swap regret of randomized response)**.** *Let $Z = (A, U)$ be a decision task with action space A and utility function U. Given a distribution $\mathcal{D}$ on $[0,1] \times \{0,1\}$ and a randomized response function $r : [0,1] \to \Delta(A)$ that maps a prediction $p \in [0,1]$ to a distribution over actions $a \in A$, we define the swap regret of r on distribution $\mathcal{D}$ as*

$$\mathsf{SR}_Z(r, \mathcal{D}) := \sup_{\sigma : A \to A} \mathbb{E}_{(p,y) \sim \mathcal{D}, a \sim r(p)} [u(\sigma(a), y) - u(a, y)].$$

We are now ready to formally state and prove the actionability of SCDL:

Theorem D.2 (SCDL is actionable)**.** *Let $\mathcal{D}$ be a distribution of $(p, y) \in [0,1] \times \{0,1\}$. Assume $\mathsf{SCDL}(\mathcal{D}) < 2/m$ for some $m \in M := \{2^1, 2^2, \ldots\}$. Then we have*

$$\sup_{Z} \mathsf{SR}_Z(r_Z^* \circ \tau_m, \mathcal{D}) = \mathsf{CDL}(\mathcal{D}_m) \leq 2\,\mathsf{SCDL}_m(\mathcal{D}) + 2/m \leq 2\,\mathsf{SCDL}(\mathcal{D}) + 2/m,$$

where the supremum is over all decision tasks $Z = (A, U)$ with a bounded utility function $U : A \times \{0,1\} \to [0,1]$.

We prove the theorem using the following lemma.

Lemma D.3. *Let $\mathcal{D}$ be a distribution on $G_m \times \{0,1\}$. We have $\mathsf{CDL}(\mathcal{D}) \le 2\,\mathsf{SCDL}_m(\mathcal{D}) + 2/m$.*

Proof. By Lemma B.2,

$$\mathsf{CDL}(\mathcal{D}) \le 2 \sup_{\mu\in[0,1]} \mathbb{E}[(q-\mu)_+\mathbb{I}[p \le \mu] + (\mu - q)_+\mathbb{I}[p > \mu]].$$

We partition the space $[0,1]$ into m intervals $B_0, \dots, B_{m-1}$ as follows. We define $B_i := [i/m, (i+1)/m)$ for $i = 0, \dots, m-2$ and $B_{m-1} = [(m-1)/m, 1]$. It remains to show that for every $i \in \{0, \dots, m-1\}$ and every $\mu \in B_i$, it holds that

$$\mathbb{E}[(q-\mu)_+\mathbb{I}[p \le \mu] + (\mu - q)_+\mathbb{I}[p \ge \mu]] \le \sum_{j=0}^{i} \pi_j (q_j - (i+1)/m)_+ + \sum_{j=i+1}^{m} \pi_j (i/m - q_j)_+ + \frac{1}{m}. \tag{10}$$

By our assumption, the predictions p lie on the grid $G_m = \{0, 1/m, \dots, 1\}$, so we have

$$\mathbb{E}[(q-\mu)_+\mathbb{I}[p \le \mu]] = \sum_{j=1}^{i} \pi_j (q_j - \mu)_+ \le \sum_{j=1}^{i} \pi_j \left((q_j - (i+1)/m)_+ + \frac{1}{m} \right).$$

The inequality above holds also in the special case where $i = m-1$ and $\mu = 1 \in B_i$, because in this case the left-hand side is zero: $(q-\mu)_+ = (q-1)_+ = 0$. Similarly,

$$\mathbb{E}[(\mu-q)_+\mathbb{I}[p > \mu]] = \sum_{j=i+1}^{m} \pi_j (\mu - q_j)_+ \le \sum_{j=i+1}^{m} \pi_j \left((i/m - q_j)_+ + \frac{1}{m} \right).$$

Combining these two inequalities with the fact that $\sum_{j=0}^{m} \pi_j = 1$ proves (10). □

Proof of Theorem D.2. By the definition of CDL (Definition B.2),

$$\sup_Z \mathsf{SR}_Z(r_Z^* \circ \tau_m, \mathcal{D}) = \mathsf{CDL}(\mathcal{D}_m). \tag{11}$$

By Lemma D.3 and Lemma D.1, we have

$$\mathsf{CDL}(\mathcal{D}_m) \le 2\,\mathsf{SCDL}_m(\mathcal{D}) + \frac{2}{m}. \tag{12}$$

By Lemma C.3, we have $m^* := m^*(\mathcal{D}) \ge m$, and

$$\mathsf{SCDL}_m(\mathcal{D}) \le \mathsf{SCDL}_{m^*}(\mathcal{D}) \le \mathsf{SCDL}(\mathcal{D}). \tag{13}$$

Combining (11), (12), and (13) completes the proof. □

Theorem D.4 (SCDL is actionable with high probability)**.** *Let $(p_1, y_1), \dots, (p_T, y_T) \in [0,1] \times \{0,1\}$ be an arbitrary sequence of prediction-outcome pairs and let $\mathcal{D}$ be the uniform distribution over these pairs. Let $m \in \{2^1, 2^2, \dots\}$ such that $\mathsf{SCDL}(\mathcal{D}) < 2/m$. For $\delta \in (0, 1/(2m))$, assume*

$$\frac{1}{m} = \Omega\left(\sqrt{\frac{\log(1/\delta)}{T}} \right).$$

For every $t = 1, \dots, T$, we independently round p_i to $p_i' \in G_m$ using the randomized rounding function τ_m in Definition D.1 and let $\mathcal{D}'$ be the uniform distribution over $(p_1', y_1), \dots, (p_T', y_T)$. Then with probability at least $1-\delta$,

$$\sup_Z \mathsf{SR}_Z(r_Z^*, \mathcal{D}') \le 2\,\mathsf{SCDL}(\mathcal{D}) + \frac{2}{m} + O\left(\sqrt{\frac{\log(1/\delta)}{T}} \cdot \log m \right),$$

where the supremum is over all decision task $Z = (A, U)$ with a bounded utility function $U : A \times \{0,1\} \to [0,1]$.

Proof. By Definition B.2, we have

$$\sup_Z \mathsf{SR}_Z(r_Z^*, \mathcal{D}') = \mathsf{CDL}(\mathcal{D}'). \tag{14}$$

By Lemma D.3, we have

$$\mathsf{CDL}(\mathcal{D}') \le 2\,\mathsf{SCDL}_m(\mathcal{D}') + \frac{2}{m}. \tag{15}$$

By Lemma G.4, with probability at least $1-\delta$,

$$\mathsf{SCDL}_m(\mathcal{D}') \le \mathsf{SCDL}_m(\mathcal{D}) + O\left(\sqrt{\frac{\log(1/\delta)}{T}} \cdot \log m\right). \tag{16}$$

By Lemma C.3, we have $m^* := m^*(\mathcal{D}) \ge m$, and

$$\mathsf{SCDL}_m(\mathcal{D}) \le \mathsf{SCDL}_{m^*}(\mathcal{D}) \le \mathsf{SCDL}(\mathcal{D}). \tag{17}$$

Combining (14), (15), (16) and (17) completes the proof. □

E Quadratic Relationship between SCDL and the Distance to Calibration

We prove the following relationship between SCDL and the lower distance to calibration DistCal (Definition B.4).

Theorem E.1 (Relationship to DistCal)**.** *For every distribution $\mathcal{D}$ on $[0,1]\times\{0,1\}$,*

$$\Omega(\mathsf{DistCal}(\mathcal{D})^2) \le \mathsf{SCDL}(\mathcal{D}) \le O(\sqrt{\mathsf{DistCal}(\mathcal{D})}).$$

Proof. For every positive integer m, we have

$$\begin{aligned}
\mathsf{DistCal}(\mathcal{D}) &\le \mathsf{DistCal}(\mathcal{D}_m) + 1/m \\
&\le \mathsf{ECE}(\mathcal{D}_m) + 1/m \\
&\le O\left(\sqrt{\mathsf{CDL}(\mathcal{D}_m)}\right) + 1/m && \text{(by Lemma B.1)} \\
&\le O\left(\sqrt{\max\{\mathsf{SCDL}_m(\mathcal{D}_m), 1/m\}}\right) && \text{(by Lemma D.3)} \\
&= O\left(\sqrt{\max\{\mathsf{SCDL}_m(\mathcal{D}), 1/m\}}\right). && \text{(by Lemma D.1)}
\end{aligned}$$

This implies $\mathsf{DistCal}(\mathcal{D}) \le O(\sqrt{\mathsf{SCDL}(\mathcal{D})})$.

Consider an arbitrary joint distribution $\mathcal{J}$ of $(p_1, p_2, y) \in [0,1]\times[0,1]\times\{0,1\}$, where the distribution of (p_1, y) is $\mathcal{D}$ and the distribution $\mathcal{D}'$ of (p_2, y) is calibrated. We have $\mathsf{SCDL}(\mathcal{D}') = 0$, so by Theorem F.1,

$$\mathsf{SCDL}(\mathcal{D}) = O(\sqrt{\mathbb{E}_{\mathcal{J}}|p_1 - p_2|}).$$

This implies $\mathsf{SCDL}(\mathcal{D}) = O(\sqrt{\mathsf{DistCal}(\mathcal{D})})$. □

Remark E.1. *The upper and lower bounds on* SCDL *in Theorem E.1 are tight up to constant factors. For every $\varepsilon \in (0, 1/4)$, Hu and Wu [22] showed a distribution $\mathcal{D}$ where* $\mathsf{DistCal}(\mathcal{D}) = \varepsilon$ *and* $\mathsf{CDL}(\mathcal{D}) = O(\varepsilon^2)$*, which implies* $\mathsf{SCDL}(\mathcal{D}) = O(\varepsilon^2)$ *by Lemma C.4. This shows that the lower bound of* SCDL *in Theorem E.1 is tight. For every $\varepsilon \in (0, 1/4)$, the distribution $\mathcal{D}_1$ in the proof of Theorem I.1 has* $\mathsf{DistCal}(\mathcal{D}_1) = O(\varepsilon)$ *and* $\mathsf{SCDL}(\mathcal{D}_1) = \Omega(\sqrt{\varepsilon})$*, proving the tightness of the upper bound.*

F Continuity of SCDL

We prove the following theorem showing that SCDL is a continuous function of the predictions.

Theorem F.1 (Continuity)**.** *Let $\mathcal{J}$ be a joint distribution of $(p_1, p_2, y) \in [0,1]\times[0,1]\times\{0,1\}$. Let $\mathcal{D}_1$ be the distribution of (p_1, y) and let $\mathcal{D}_2$ be the distribution of (p_2, y). We have*

$$|\mathsf{SCDL}(\mathcal{D}_1)^2 - \mathsf{SCDL}(\mathcal{D}_2)^2| \le O(\mathbb{E}_{\mathcal{J}}|p_1 - p_2|), \quad \text{and}$$

$$|\mathsf{SCDL}(\mathcal{D}_1) - \mathsf{SCDL}(\mathcal{D}_2)| \le O\left(\sqrt{\mathbb{E}_{\mathcal{J}}|p_1 - p_2|}\right).$$

We need the following helper lemma:

Lemma F.2. *Let $\mathcal{J}$ be a joint distribution of $(p_1, p_2, y) \in [0,1] \times [0,1] \times \{0,1\}$. Let $\mathcal{D}_1$ be the distribution of (p_1, y) and let $\mathcal{D}_2$ be the distribution of (p_2, y). We have*

$$|\mathsf{SCDL}_m(\mathcal{D}_1) - \mathsf{SCDL}_m(\mathcal{D}_2)| \le O(m \cdot \mathbb{E}_{\mathcal{J}}|p_1 - p_2|) \quad \textit{for every } m = 0, 1, \ldots.$$

Proof. It suffices to show that for every $i = 0, \ldots, m$,

$$|\mathsf{SCFDL}_{m,i}(\mathcal{D}_1) - \mathsf{SCFDL}_{m,i}(\mathcal{D}_2)| \le O(m \cdot \mathbb{E}_{\mathcal{J}}|p_1 - p_2|).$$

We first show that for any two predictions $p, p' \in [0,1]$, it holds that

$$\sum_{i=0}^{m} |w_i(p) - w_i(p')| \le O(m|p - p'|).$$

This is because for all but at most 4 choices of i we have $w_i(p) = w_i(p') = 0$, where the 4 choices are the two neighboring bins to p and the two neighboring bins to p'. For these 4 choices of i, we have $|w_i(p) - w_i(p')| \le O(m|p - p'|)$.

Now we have

$$\begin{aligned}
\sum_{i=0}^{m} |\pi_i - \pi'_i| &= \sum_{i=0}^{m} |\mathbb{E}_{\mathcal{J}}(w_i(p) - w_i(p'))| \\
&\le \mathbb{E}_{\mathcal{J}} \sum_{i=0}^{m} |w_i(p) - w_i(p')| \\
&\le O(m \cdot \mathbb{E}_{\mathcal{J}}|p - p'|).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sum_{i=0}^{m} |\pi_i q_i - \pi'_i q'_i| &= \sum_{i=0}^{m} |\mathbb{E}_{\mathcal{J}}(w_i(p)y - w_i(p')y)| \\
&\le \mathbb{E}_{\mathcal{J}} \sum_{i=0}^{m} |w_i(p) - w_i(p')|y \\
&\le O(m \cdot \mathbb{E}_{\mathcal{J}}|p - p'|).
\end{aligned}$$

By the definition of SCFDL and the triangle inequality,

$$\begin{aligned}
&|\mathsf{SCFDL}_{m,i}(\mathcal{D}_1) - \mathsf{SCFDL}_{m,i}(\mathcal{D}_2)| \\
\le &\sum_{j \le i} |(\pi_j q_j - \pi_j(i+1)/m)_+ - (\pi'_j q'_j - \pi'_j(i+1)/m)_+| \\
&+ \sum_{j > i} |(\pi_j i/m - \pi_j q_j)_+ + (\pi'_j i/m - \pi'_j q'_j)_+| \\
\le &\sum_{j=0}^{m} |\pi_j q_j - \pi'_j q'_j| + \sum_{j=0}^{m} |\pi_j - \pi'_j| \\
\le &O(m \cdot \mathbb{E}_{\mathcal{J}}|p - p'|).
\end{aligned}$$

□

Proof of Theorem F.1. Assume without loss of generality that $\mathbb{E}_{\mathcal{J}}|p_1 - p_2| > 0$ and $\mathsf{SCDL}(\mathcal{D}_1) \le \mathsf{SCDL}(\mathcal{D}_2)$. Define $m^* := m^*(\mathcal{D}_1)$. By Lemma F.2, for an absolute constant $C > 0$ and every $m = 0, 1, \ldots$ satisfying $m \le m^*$,

$$\mathsf{SCDL}_m(\mathcal{D}_2) \le \mathsf{SCDL}(\mathcal{D}_1) + Cm \cdot \mathbb{E}_{\mathcal{J}}|p_1 - p_2|. \tag{18}$$

Choose m to be smallest number among $2^1, 2^2, \ldots$ satisfying

$$\mathsf{SCDL}(\mathcal{D}_1) + Cm \cdot \mathbb{E}_{\mathcal{J}}|p_1 - p_2| \ge \frac{1}{m}. \tag{19}$$

We have $m \le m^*$ because $\mathsf{SCDL}(\mathcal{D}_1) \ge 1/m^*$ (Lemma C.3). By (18),

$$\mathsf{SCDL}(\mathcal{D}_2) \le \max\{\mathsf{SCDL}_m(\mathcal{D}_2), 1/m\} \le \mathsf{SCDL}(\mathcal{D}_1) + Cm \cdot \mathbb{E}_{\mathcal{J}}|p_1 - p_2|.$$

Therefore,

$$\begin{aligned}
\mathsf{SCDL}(\mathcal{D}_2)^2 - \mathsf{SCDL}(\mathcal{D}_1)^2 &= (\mathsf{SCDL}(\mathcal{D}_2) - \mathsf{SCDL}(\mathcal{D}_1))(\mathsf{SCDL}(\mathcal{D}_1) + \mathsf{SCDL}(\mathcal{D}_2)) \\
&\le Cm \cdot \mathbb{E}_{\mathcal{J}}|p_1 - p_2| \ (2 \cdot \mathsf{SCDL}(\mathcal{D}_1) + Cm \cdot \mathbb{E}_{\mathcal{J}}|p_1 - p_2|) \\
&\le Cm \cdot \mathbb{E}_{\mathcal{J}}|p_1 - p_2| \cdot O(1/m) \\
&\qquad \text{(because } m \text{ is the smallest power of 2 satisfying (19))} \\
&\le O(\mathbb{E}_{\mathcal{J}}|p_1 - p_2|).
\end{aligned}$$

Finally,

$$\begin{aligned}
(\mathsf{SCDL}(\mathcal{D}_2) - \mathsf{SCDL}(\mathcal{D}_1))^2 &\le (\mathsf{SCDL}(\mathcal{D}_2) - \mathsf{SCDL}(\mathcal{D}_1))(\mathsf{SCDL}(\mathcal{D}_2) + \mathsf{SCDL}(\mathcal{D}_1)) \\
&= (\mathsf{SCDL}(\mathcal{D}_2)^2 - \mathsf{SCDL}(\mathcal{D}_1)^2) \\
&\le O(\mathbb{E}_{\mathcal{J}}|p_1 - p_2|).
\end{aligned}$$

□

G Testability of SCDL

We formally state and prove the testability of SCDL.

Theorem G.1 (SCDL is testable)**.** *Let S be the uniform distribution over $T \ge 2$ data points $(p_1, y_1), \ldots, (p_T, y_T)$ drawn i.i.d. from a distribution $\mathcal{D}$ on $[0,1] \times \{0,1\}$. For every $\delta \in (e^{-T/2}, 1/T)$, with probability at least $1 - \delta$ we have*

$$|\mathsf{SCDL}(S) - \mathsf{SCDL}(\mathcal{D})| \le O\left(\sqrt{\frac{\log(1/\delta)}{T}} \log \frac{T}{\ln(1/\delta)}\right). \tag{20}$$

Moreover, when $\mathsf{SCDL}(\mathcal{D}) = 0$ the following improved bound holds: for every $\delta \in (e^{-T/2}, 1/T)$, with probability at least $1 - \delta$ we have

$$\mathsf{SCDL}(S) \le O\left(\sqrt{\frac{\log(1/\delta)}{T} \log \frac{T}{\ln(1/\delta)}}\right). \tag{21}$$

Our proof of Theorem G.1 relies on the following lemma:

Lemma G.2. *Let S be the uniform distribution over T data points $(p_1, y_1), \ldots, (p_T, y_T)$ drawn i.i.d. from a distribution $\mathcal{D}$ on $[0,1] \times \{0,1\}$. For every positive integer m and every $i = 0, \ldots, m$, for every $\delta \in (0, 1/(2m))$, with probability at least $1 - \delta$ we have*

$$\begin{aligned}
&|\mathsf{SCFDL}_{m,i}(S) - \mathsf{SCFDL}_{m,i}(\mathcal{D})| \\
&\le O\left(\sqrt{\frac{\log(1/\delta)}{T}} + (\log m)\sqrt{\frac{m \cdot \mathsf{SCDL}_m(\mathcal{D}) \log(1/\delta)}{T}} + \frac{m(\log m)\log(1/\delta)}{T}\right).
\end{aligned}$$

We prove Lemma G.2 using Lemma G.3 below.

Lemma G.3. *Let $\mathcal{D}$ be a distribution on $[0,1] \times \{0,1\}$. For every positive integer m, every $i = 0, \ldots, m$, define π_i and q_i as in Lemma C.1. For every $\alpha > 0$, we have*

$$\sum_{j<i} \mathbb{I}[q_j > i/m - \alpha/\sqrt{\pi_j}]\sqrt{\pi_j} \le 3 + O(\log m)(\alpha m + \sqrt{m \cdot \mathsf{SCDL}_m(\mathcal{D})}). \tag{22}$$

Proof. When $m = 1$, the index j can only take two possible values: 0 and 1. In this case, the left-hand side of (22) is at most 2, and the lemma holds trivially. We assume $m \ge 2$ from now on.

For $k = 0, 1, 2, 3, \ldots$, define $S_k := \{j \in \{0, \ldots, m\} : 2^{-k-1} < \pi_j \le 2^{-k}\}$. Define

$$\begin{aligned}
i_k &:= i - \lceil 2^{(k+1)/2} \alpha m \rceil \le i - 1, \\
S'_k &:= \{j \in S_k : i_k - 1 \le j \le i - 3\},
\end{aligned}$$

$$S_k'' := \{j \in S_k : j \le i_k - 2, q_j > i_k/m\}.$$

We prove the following two inequalities:

$$|S_k'| \le O(2^{k/2}\alpha m), \tag{23}$$
$$|S_k''| \le O(2^{k/2}\sqrt{m \cdot \mathsf{SCDL}_m(\mathcal{D})}). \tag{24}$$

By the definition of S_k' and i_k, we have $|S_k'| \le i - i_k - 1 = \lceil 2^{(k+1)/2}\alpha m\rceil - 1 \le O(2^{k/2}\alpha m)$. This proves (23). Define $i_k' := i_k - \lceil |S_k''|/2\rceil - 1$. There are at least $|S_k''|/2$ different $j \in S_k''$ satisfying $j \le i_k'$. Therefore,

$$\mathsf{SCDL}_m(\mathcal{D}) \ge \mathsf{SCFDL}_{m,i_k'}(\mathcal{D}) \ge 2^{-k-1}(|S_k''|/2)(i_k/m - (i_k'+1)/m) \ge 2^{-k-1}(|S_k''|/2)^2/m.$$

Rearranging the inequality above proves (24). The lemma is proved by the following calculation:

$$\begin{aligned}
&\sum_{j<i} \mathbb{I}[q_j > i/m - \alpha/\sqrt{\pi_j}]\sqrt{\pi_j} \\
&\le \sqrt{\pi_{i-1}} + \sqrt{\pi_{i-2}} \\
&\quad + \sum_{j\le i, \pi_j \le 1/m^2} \sqrt{\pi_j} + \sum_{k=0}^{\lceil \log_2(m^2)\rceil} \sum_{j \le i-3} \mathbb{I}[2^{-k-1} < \pi_j \le 2^{-k}, q_j > i/m - \alpha/\sqrt{\pi_j}]\sqrt{\pi_j} \\
&\le 1 + 1 + 1 + \sum_{k=0}^{\lceil \log_2(m^2)\rceil} \sum_{j \le i-3} \mathbb{I}[2^{-k-1} < \pi_j \le 2^{-k}, q_j > i_k/m]\sqrt{\pi_j} \\
&\le 3 + \sum_{k=0}^{\lceil \log_2(m^2)\rceil} \sum_{j \le i-3} \mathbb{I}[2^{-k-1} < \pi_j \le 2^{-k}, q_j > i_k/m] \cdot 2^{-k/2} && \text{(because } \pi_j \le 2^{-k}\text{)} \\
&\le 3 + \sum_{k=0}^{\lceil \log_2(m^2)\rceil} 2^{-k/2}|S_k' \cup S_k''| \\
&\le 3 + O(\log m)(\alpha m + \sqrt{m \cdot \mathsf{SCDL}_m(\mathcal{D})}).
\end{aligned}$$

□

Proof of Lemma G.2. For every $j = 0, \dots, m$, define π_j and q_j as in Lemma C.1. Define

$$\hat{\pi}_j := \mathbb{E}_{(p,y)\sim S}[w_i(p)] = \frac{1}{T}\sum_{t=1}^{T} w_i(p_t)$$
$$\hat{q}_j := \frac{1}{\hat{\pi}_j}\mathbb{E}_{(p,y)\sim S}[w_i(p)y] = \frac{1}{T\hat{\pi}_j}\sum_{t=1}^{T} w_i(p_t)y_t.$$

By Bernstein's inequality and the union bound, with probability at least $1-\delta$, for every $j = 0, \dots, m$,

$$|\hat{\pi}_j - \pi_j| \le O\left(\sqrt{\frac{\pi_j \log(1/\delta)}{T}} + \frac{\log(1/\delta)}{T}\right), \tag{25}$$
$$|\hat{\pi}_j\hat{q}_j - \pi_j q_j| \le O\left(\sqrt{\frac{\pi_j \log(1/\delta)}{T}} + \frac{\log(1/\delta)}{T}\right). \tag{26}$$

We assume that (25) and (26) hold. For some absolute constant $C > 0$, as long as $\pi_j \ge \frac{C\log(1/\delta)}{T}$, we have

$$|\hat{\pi}_j - \pi_j| \le O\left(\sqrt{\frac{\pi_j \log(1/\delta)}{T}}\right), \tag{27}$$

$$|\hat{\pi}_j\hat{q}_j - \pi_j q_j| \le O\left(\sqrt{\frac{\pi_j \log(1/\delta)}{T}}\right), \quad \text{and} \tag{28}$$

$$|\hat{q}_j - q_j| \le \frac{1}{\pi_j}(|\hat{\pi}_j\hat{q}_j - \pi_j q_j| + |\hat{\pi}_j - \pi_j|\hat{q}_j) \le O\left(\sqrt{\frac{\log(1/\delta)}{T\pi_j}}\right). \tag{29}$$

The last inequality implies $|\hat{q}_j - q_j| \le \alpha/\sqrt{\pi_j}$ for some $\alpha = O\left(\sqrt{\frac{\log(1/\delta)}{T}}\right)$.

Define H_- as the set of indices $j \in \{0, \dots, i\}$ satisfying $\max\{q_j, \hat{q}_j\} > (i+1)/m$. Define H_+ as the set of indices $j \in \{i+1, \dots, m\}$ satisfying $\min\{q_j, \hat{q}_j\} < i/m$. By the triangle inequality,

$$\begin{aligned}
&|\mathsf{SCFDL}_{m,i}(S) - \mathsf{SCFDL}_{m,i}(\mathcal{D})| \\
&\le \sum_{j\le i} |\hat{\pi}_j(\hat{q}_j - (i+1)/m)_+ - \pi_j(q_j - (i+1)/m)_+| + \sum_{j>i} |\hat{\pi}_j(i/m - \hat{q}_j)_+ - \pi_j(i/m - q_j)_+| \\
&\le \sum_{j\in H_-} |\hat{\pi}_j(\hat{q}_j - (i+1)/m) - \pi_j(q_j - (i+1)/m)| + \sum_{j\in H_+} |\hat{\pi}_j(i/m - \hat{q}_j) - \pi_j(i/m - q_j)| \\
&\le \sum_{j\in H_-} |\hat{\pi}_j\hat{q}_j - \pi_j q_j| + \sum_{j\in H_-} |\pi_j - \hat{\pi}_j| + \sum_{j\in H_+} \pi_j|q_j - \hat{q}_j| + \sum_{j\in H_+} |\pi_j - \hat{\pi}_j|.
\end{aligned} \tag{30}$$

Define L as the set of indices $j = 0, \dots, m$ satisfying $\pi_j < \frac{C\log(1/\delta)}{T}$. By (27), we have $\hat{\pi}_j = O(\frac{\log(1/\delta)}{T})$. Therefore,

$$\sum_{j\in L} |\hat{\pi}_j\hat{q}_j - \pi_j q_j| + \sum_{j\in L} |\pi_j - \hat{\pi}_j| \le 2|L| \cdot O\left(\frac{\log(1/\delta)}{T}\right) = O\left(\frac{m\log(1/\delta)}{T}\right). \tag{31}$$

Moreover, by (28) and (29),

$$\begin{aligned}
&\sum_{j\in H_-\setminus L} |\hat{\pi}_j\hat{q}_j - \pi_j q_j| + \sum_{j\in H_-\setminus L} |\hat{\pi}_j - \pi_j| \\
&\le O\left(\sqrt{\frac{\log(1/\delta)}{T}}\right)\left(\sum_{j\in H_-\setminus L} \sqrt{\pi_j}\right) \\
&\le O\left(\sqrt{\frac{\log(1/\delta)}{T}}\right) \cdot (3 + O(\log m)(\alpha m + \sqrt{m \cdot \mathsf{SCDL}_m(\mathcal{D})})) \qquad \text{(by Lemma G.3)} \\
&= O\left(\sqrt{\frac{\log(1/\delta)}{T}} + (\log m)\sqrt{\frac{m \cdot \mathsf{SCDL}_m(\mathcal{D})\log(1/\delta)}{T}} + \frac{m(\log m)\log(1/\delta)}{T}\right).
\end{aligned} \tag{32}$$

Similarly, we have

$$\begin{aligned}
&\sum_{j\in H_+\setminus L} |\hat{\pi}_j\hat{q}_j - \pi_j q_j| + \sum_{j\in H_+\setminus L} |\hat{\pi}_j - \pi_j| \\
&= O\left(\sqrt{\frac{\log(1/\delta)}{T}} + (\log m)\sqrt{\frac{m \cdot \mathsf{SCDL}_m(\mathcal{D})\log(1/\delta)}{T}} + \frac{m(\log m)\log(1/\delta)}{T}\right).
\end{aligned} \tag{33}$$

Combining (31), (32), and (33) and plugging the result into (30) completes the proof. □

The same proof of Lemma G.2 establishes the following lemma, which we used in Appendix D to prove Theorem D.4.

Lemma G.4. *Let $(p_1, y_1), \dots, (p_T, y_T) \in [0,1] \times \{0,1\}$ be an arbitrary sequence of prediction-outcome pairs and let $\mathcal{D}$ be the uniform distribution over these pairs. Let m be a positive integer. For every $t = 1, \dots, T$, we independently round p_i to $p'_i \in G_m$ using the randomized rounding function τ_m in Definition D.1 and let $\mathcal{D}'$ be the uniform distribution over $(p'_1, y_1), \dots, (p'_T, y_T)$. Then for every $\delta \in (0, 1/(2m))$, with probability at least $1 - \delta$ we have*

$$|\mathsf{SCDL}_m(\mathcal{D}') - \mathsf{SCDL}_m(\mathcal{D})|$$

$$\leq O\left(\sqrt{\frac{\log(1/\delta)}{T}} + (\log m)\sqrt{\frac{m \cdot \mathsf{SCDL}_m(\mathcal{D})\log(1/\delta)}{T}} + \frac{m(\log m)\log(1/\delta)}{T}\right).$$

Proof of Theorem G.1. It holds trivially that $|\mathsf{SCDL}(S) - \mathsf{SCDL}(\mathcal{D})| \leq 1$. We can thus assume without loss of generality that $T/\log(1/\delta) \geq 100C_0$ for an absolute constant $C_0 > 0$ which we will determine later. Let $m_{\max}$ be the largest number in $M := \{2^1, 2^2, \dots, \}$ such that $m_{\max} \log m_{\max} \leq \sqrt{\frac{T}{C_0 \log(1/\delta)}}$.

Define $m^* := m^*(\mathcal{D}) \in M \cup \{+\infty\}$. We divide the analysis into two cases: $m^* < m_{\max}$ and $m^* \geq m_{\max}$.

When $m^* < m_{\max}$, by Lemma G.2 and the union bound, with probability at least $1 - \delta$ the following two conditions hold simultaneously:

$$\begin{aligned}
&|\mathsf{SCDL}_{m^*}(S) - \mathsf{SCDL}_{m^*}(\mathcal{D})| \\
&\leq O\left(\sqrt{\frac{\log(1/\delta)}{T}} + (\log m^*)\sqrt{\frac{m^* \cdot \mathsf{SCDL}_{m^*}(\mathcal{D})\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\log(1/\delta)}{T}\right) \\
&\leq O\left((\log m^*)\sqrt{\frac{\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\log(1/\delta)}{T}\right), \qquad (34)
\end{aligned}$$

and

$$\begin{aligned}
&|\mathsf{SCDL}_{2m^*}(S) - \mathsf{SCDL}_{2m^*}(\mathcal{D})| \\
&\leq O\left(\sqrt{\frac{\log(1/\delta)}{T}} + (\log m^*)\sqrt{\frac{m^* \cdot \mathsf{SCDL}_{2m^*}(\mathcal{D})\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\log(1/\delta)}{T}\right) \\
&\leq O\left((\log m^*)\sqrt{\frac{m^* \cdot \mathsf{SCDL}_{2m^*}(\mathcal{D})\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\log(1/\delta)}{T}\right), \qquad (35)
\end{aligned}$$

where we used $m^* \cdot \mathsf{SCDL}_{m^*}(\mathcal{D}) < 2$ and $m^* \cdot \mathsf{SCDL}_{2m^*}(\mathcal{D}) \geq 1$ implied by Lemma C.3 to simplify the expressions. We assume (34) and (35) hold from now on.

By (35), there exists an absolute constant $C > 0$ such that

$$\begin{aligned}
&\mathsf{SCDL}_{2m^*}(S) \\
&\geq \mathsf{SCDL}_{2m^*}(\mathcal{D}) - C\left((\log m^*)\sqrt{\frac{m^* \cdot \mathsf{SCDL}_{2m^*}(\mathcal{D})\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\log(1/\delta)}{T}\right). \qquad (36)
\end{aligned}$$

By Lemma C.3, $\mathsf{SCDL}_{2m^*}(\mathcal{D}) \geq \mathsf{SCDL}(\mathcal{D}) \geq 1/m^*$. By our choice of $m_{\max}$, we have

$$m^* \log m^* \leq m_{\max} \log m_{\max} \leq \sqrt{\frac{T}{C_0 \log(1/\delta)}}. \qquad (37)$$

Choosing C_0 to be a sufficiently large constant, we can ensure that the right-hand side of (36) is an increasing function of $\mathsf{SCDL}_{2m^*}(\mathcal{D}) \in [1/m^*, 1]$. Therefore, (36) implies

$$\begin{aligned}
&\mathsf{SCDL}_{2m^*}(S) \\
&\geq \mathsf{SCDL}(\mathcal{D}) - C\left((\log m^*)\sqrt{\frac{m^* \cdot \mathsf{SCDL}(\mathcal{D})\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\log(1/\delta)}{T}\right) \\
&\geq \mathsf{SCDL}(\mathcal{D}) - O\left((\log m^*)\sqrt{\frac{\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\ln(1/\delta)}{T}\right), \qquad (38)
\end{aligned}$$

where the last inequality in (38) holds because $m^* \cdot \mathsf{SCDL}(\mathcal{D}) < 2$.

Let us establish an upper bound on $\mathsf{SCDL}(S)$:

$$\mathsf{SCDL}(S) \leq \max\{\mathsf{SCDL}_{m^*}(S), 1/m^*\}$$

$$\leq \max\{\mathsf{SCDL}_{m^*}(\mathcal{D}), 1/m^*\} + |\mathsf{SCDL}_{m^*}(S) - \mathsf{SCDL}_{m^*}(\mathcal{D})|$$
$$= \mathsf{SCDL}(\mathcal{D}) + |\mathsf{SCDL}_{m^*}(S) - \mathsf{SCDL}_{m^*}(\mathcal{D})|.$$

Plugging (34) into the inequality above, we get

$$\mathsf{SCDL}(S) \leq \mathsf{SCDL}(\mathcal{D}) + O\left((\log m^*)\sqrt{\frac{\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\log(1/\delta)}{T}\right). \tag{39}$$

We now turn to proving a lower bound for $\mathsf{SCDL}(S)$:

$$\begin{aligned}\mathsf{SCDL}(S) = \min\Big\{ &\inf_{m<m^*} \max\{\mathsf{SCDL}_m(S), 1/m\}, \\ &\max\{\mathsf{SCDL}_{m^*}(S), 1/m^*\}, \\ &\inf_{m>m^*} \max\{\mathsf{SCDL}_m(S), 1/m\}\Big\} \\ \geq \min\{&2/m^*, \max\{\mathsf{SCDL}_{m^*}(S), 1/m^*\}, \mathsf{SCDL}_{2m^*}(S)\}.\end{aligned} \tag{40}$$

The first two terms in the minimum above can be lower-bounded as follows using Lemma C.3:

$$\begin{aligned}2/m^* &> \mathsf{SCDL}(\mathcal{D}), \\ \max\{\mathsf{SCDL}_{m^*}(S), 1/m^*\} &\geq \max\{\mathsf{SCDL}_{m^*}(\mathcal{D}), 1/m^*\} - |\mathsf{SCDL}_{m^*}(S) - \mathsf{SCDL}_{m^*}(\mathcal{D})| \\ &= \mathsf{SCDL}(\mathcal{D}) - |\mathsf{SCDL}_{m^*}(S) - \mathsf{SCDL}_{m^*}(\mathcal{D})|.\end{aligned}$$

The third term $\mathsf{SCDL}_{2m^*}(S)$ can be lower-bounded by (38). Plugging these into (40) and using (34), we get

$$\mathsf{SCDL}(S) \geq \mathsf{SCDL}(\mathcal{D}) - O\left((\log m^*)\sqrt{\frac{\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\log(1/\delta)}{T}\right). \tag{41}$$

Combining (39) and (41), we get

$$|\mathsf{SCDL}(S) - \mathsf{SCDL}(\mathcal{D})| = O\left((\log m^*)\sqrt{\frac{\log(1/\delta)}{T}} + \frac{m^*(\log m^*)\log(1/\delta)}{T}\right).$$

Combining this with (37) proves (20).

When $m^* \geq m_{\max}$, Lemma C.3 implies

$$\mathsf{SCDL}_{m_{\max}}(\mathcal{D}) \leq \mathsf{SCDL}_{m^*}(\mathcal{D}) \leq \mathsf{SCDL}(\mathcal{D}) < 2/m_{\max}. \tag{42}$$

By Lemma G.2 and the union bound, with probability at least $1-\delta$ it holds that

$$|\mathsf{SCDL}_{m_{\max}}(S) - \mathsf{SCDL}_{m_{\max}}(\mathcal{D})| = O\left((\log m_{\max})\sqrt{\frac{\log(1/\delta)}{T}} + \frac{m_{\max}(\log m_{\max})\log(1/\delta)}{T}\right). \tag{43}$$

By (42),

$$\begin{aligned}\mathsf{SCDL}(S) &\leq \max\{\mathsf{SCDL}_{m_{\max}}(S), 1/m_{\max}\} \\ &\leq \max\{\mathsf{SCDL}_{m_{\max}}(\mathcal{D}), 1/m_{\max}\} + |\mathsf{SCDL}_{m_{\max}}(S) - \mathsf{SCDL}_{m_{\max}}(\mathcal{D})| \\ &\leq 2/m_{\max} + |\mathsf{SCDL}_{m_{\max}}(S) - \mathsf{SCDL}_{m_{\max}}(\mathcal{D})|.\end{aligned} \tag{44}$$

Plugging (43) into (44) and using $\mathsf{SCDL}(\mathcal{D}) < 2/m_{\max}$, we have

$$|\mathsf{SCDL}(S) - \mathsf{SCDL}(\mathcal{D})| \leq O\left(\frac{1}{m_{\max}} + (\log m_{\max})\sqrt{\frac{\log(1/\delta)}{T}} + \frac{m_{\max}(\log m_{\max})\log(1/\delta)}{T}\right).$$

This proves (20) by our choice of $m_{\max}$.

When $\mathsf{SCDL}(\mathcal{D}) = 0$, we choose $m_{\max}$ as the largest number in M such that

$$m_{\max}\sqrt{\log m_{\max}} \leq \sqrt{\frac{T}{C_0 \log(1/\delta)}}.$$

By Lemma G.2 and the union bound, with probability at least $1-\delta$ we have

$$\mathsf{SCDL}_{m_{\max}}(S) = O\left(\sqrt{\frac{\log(1/\delta)}{T}} + \frac{m_{\max}(\log m_{\max})\log(1/\delta)}{T}\right).$$

Therefore,

$$\begin{aligned}\mathsf{SCDL}(S) &\le \max\{\mathsf{SCDL}_{m_{\max}}(S), 1/m_{\max}\} \\ &= O\left(\frac{1}{m_{\max}} + \sqrt{\frac{\log(1/\delta)}{T}} + \frac{m_{\max}(\log m_{\max})\log(1/\delta)}{T}\right).\end{aligned}$$

This proves (21) by our choice of $m_{\max}$. □

H Lower bound

In this section, we prove an $\Omega(T^{-1/2})$ lower bound for the estimation error $\varepsilon(T)$ in Question 1, showing that our upper bound is tight up to a logarithmic factor.

Theorem H.1. *Any calibration measure CAL_T solving Question 1 requires $\varepsilon(T) \ge \Omega(T^{-1/2})$.*

We use the following lemma to prove Theorem H.1.

Lemma H.2. *There exists a perfectly calibrated distribution $\mathcal{D}$ of $(p, y) \in [0,1] \times \{0,1\}$ and a decision task $Z = (A, U)$ with the following properties:*

1. *The utility function U is bounded between 0 and 1, i.e., $U : A \times \{0,1\} \to [0,1]$.*
2. *Let S be the uniform distribution on T i.i.d. data points drawn from $\mathcal{D}$. For every positive integer T and every randomized response function $r : [0,1] \to \Delta(A)$, with probability at least $1/3$ over the randomness in S,*

$$\mathsf{SR}_Z(r, S) = \Omega(T^{-1/2}).$$

Proof. Let $\mathcal{D}$ be the uniform distribution on $(1/2, 0)$ and $(1/2, 1)$. Let U be the simple decision task with binary actions $A = \{0,1\}$ and a 0-1 utility function $U(a, y) = \mathbb{I}[a = y]$. Consider the action distribution $r(1/2)$ produced by an arbitrary randomized response function r. Since there are only two actions $a = 0, 1$, $r(1/2)$ must put probability mass at least $1/2$ on one of the two actions. We assume without loss of generality that $r(1/2)$ puts at least probability $1/2$ on action $a = 0$.

On a random sample of T i.i.d. points $(p_1, y_1), \ldots, (p_T, y_T)$ drawn from $\mathcal{D}$, we have $p_1 = \cdots = p_T = 1/2$. Also, $y_1, \ldots, y_T$ are drawn independently from $\mathsf{Ber}(1/2)$, so with probability at least $1/3$, we have

$$\frac{1}{T}\sum_{t=1}^{T} y_t \ge 1/2 + \Omega(T^{-1/2}). \tag{45}$$

It remains to show that whenever (45) holds, we have $\mathsf{SR}_Z(r, S) = \Omega(T^{-1/2})$. When (45) holds, for (p, y) drawn from S, we have $p = 1/2$ deterministically and y is drawn from a Bernoulli distribution with mean $1/2 + \Omega(T^{-1/2})$. For this distribution of y, the best action to take is $a = 1$. A probability q of taking the wrong action $a = 0$ incurs swap regret $\Omega(qT^{-1/2})$. Since $r(1/2)$ puts probability at least $1/2$ on action $a = 0$, it incurs swap regret $\Omega(T^{-1/2})$. □

Proof of Theorem H.1. Let CAL_T be a solution to Question 1. Consider the distribution $\mathcal{D}$ and the decision task Z from Lemma H.2. The second requirement of Question 1 implies $\mathbb{E}[\mathsf{CAL}_T(S_T)] \le \varepsilon(T)$, where S_T is the uniform distribution over $(p_1, y_1), \ldots, (p_T, y_T)$ drawn i.i.d. from $\mathcal{D}$. By Markov's inequality, $\Pr[\mathsf{CAL}_T(S_T) \le 4\varepsilon(T)] \ge 3/4$. Setting $\alpha = 4\varepsilon(T)$ in the first requirement of Question 1, we have $\Pr[\mathsf{SR}_Z(r_{Z,\alpha}, S_T) \le 4\varepsilon(T)] \ge 3/4$. By Lemma H.2, $\Pr[\mathsf{SR}_Z(r_{Z,\alpha}, S_T) \ge \Omega(T^{-1/2})] \ge 1/3$. Combining the last two inequalities, we get $\varepsilon(T) \ge \Omega(T^{-1/2})$. □

I Quadratic Gap between Smooth Calibration Error and Swap Regret

Theorem I.1 ([20]). *For every $\varepsilon \in [0, 1/8)$ there exist two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ of $(p, y) \in [0, 1] \times \{0, 1\}$ and a decision task $Z = (A, U)$ such that*

1. $\mathsf{smCE}(\mathcal{D}_1) = \mathsf{smCE}(\mathcal{D}_2) = \varepsilon$,
2. *the utility function U is bounded between 0 and 1, i.e., $U : A \times \{0, 1\} \to [0, 1]$, and*
3. *for every randomized response function $r : [0, 1] \to \Delta(A)$,*

$$\max\{\mathsf{SR}_Z(r, \mathcal{D}_1), \mathsf{SR}_Z(r, \mathcal{D}_2)\} \geq \Omega(\sqrt{\varepsilon}).$$

Proof. Let $\mathcal{D}_1$ be the distribution of $(p, y) \in [0, 1] \times \{0, 1\}$ where p is uniform from $\{1/2 - \sqrt{\varepsilon}, 1/2 + \sqrt{\varepsilon}\}$ and y is drawn independently and uniformly from $\{0, 1\}$. Let $\mathcal{D}_2$ be the distribution of (p, y) where $p = 1/2 - \sqrt{\varepsilon}$ deterministically and $y \sim \mathsf{Ber}(1/2 - \sqrt{\varepsilon} - c\varepsilon)$ for a parameter $c > 0$. Any $c = O(1)$ ensures that both $\mathcal{D}_1$ and $\mathcal{D}_2$ have smooth calibration error $O(\varepsilon)$, and we can choose c appropriately to make the two distributions have the same smooth calibration error. Consider a simple decision task with binary actions $A = \{0, 1\}$ and utility function

$$u(a, y) = (1 - \sqrt{\varepsilon})\mathbb{I}[a = 0, y = 0] + \mathbb{I}[a = 1, y = 1].$$

On $\mathcal{D}_1$, the optimal action to take is always $a = 1$, regardless of the prediction p. On $\mathcal{D}_2$, the optimal action to take is always $a = 0$. A probability q of taking the wrong action incurs $\Omega(q\sqrt{\varepsilon})$ expected swap regret, and in fact $\Omega(q\sqrt{\varepsilon})$ external regret as well. Since both $\mathcal{D}_1$ and $\mathcal{D}_2$ put a probability mass of at least $1/2$ on the same prediction $1/2 - \sqrt{\varepsilon}$, any post-processing, even a randomized one, will lead to a probability of at least $1/4$ taking the wrong action on either $\mathcal{D}_1$ or $\mathcal{D}_2$. □

J Power-2/3 Gap between Cutoff Calibration Error and Swap Regret

Theorem J.1. *Let $Z = (A, U)$ be the decision task where $A = [0, 1]$ and $U(a, y) = 1 - (a - y)^2$. For every $\varepsilon \in (0, 10^{-100})$ and every randomized response function $r : [0, 1] \to \Delta(A)$, there exists a distribution $\mathcal{D}$ with $\mathsf{Cutoff}(\mathcal{D}) = \varepsilon$ and $\mathsf{SR}_Z(r, \mathcal{D}) \geq \Omega(\varepsilon^{2/3})$.*

The proof of the this theorem as well as the exponent $2/3$ was discovered by ChatGPT Pro. We independently verified that the proof is correct and present it below. We first prove the following helper lemma:

Lemma J.2. *Let Z be the decision task from Theorem J.1. Let $\mathcal{D}$ be a distribution on $[0, 1] \times \{0, 1\}$ and let $r : [0, 1] \to \Delta([0, 1])$ be a randomized response function. Let Γ be the distribution of $(a, y) \in [0, 1] \times \{0, 1\}$ where we first draw (p, y) from $\mathcal{D}$ and then sample $a \sim r(p)$. We have*

$$\mathsf{SR}_Z(r, \mathcal{D}) = \mathbb{E}_\Gamma[(a - \mathbb{E}_\Gamma[y|a])^2].$$

Moreover, for every function $\varphi : [0, 1] \to \mathbb{R}$, it holds that

$$\mathsf{SR}_Z(r, \mathcal{D})\mathbb{E}_\Gamma[\varphi(a)^2] \geq \mathbb{E}_\Gamma[(y - a)\varphi(a)]^2. \tag{46}$$

Proof. By the definition of swap regret,

$$\mathsf{SR}_Z(r, \mathcal{D}) = \sup_{\sigma:[0,1]\to[0,1]} \mathbb{E}_{(a,y)\sim\Gamma}[(a - y)^2 - (\sigma(a) - y)^2]. \tag{47}$$

We can simply the expression using the following calculation:

$$\begin{aligned}\mathbb{E}[(a - y)^2] &= \mathbb{E}[(a - \mathbb{E}[y|a])^2] + \mathbb{E}[(y - \mathbb{E}[y|a])^2] - 2\,\mathbb{E}[(a - \mathbb{E}[y|a])(y - \mathbb{E}[y|a])] \\ &= \mathbb{E}[(a - \mathbb{E}[y|a])^2] + \mathbb{E}[(y - \mathbb{E}[y|a])^2].\end{aligned}$$

Similarly, $\mathbb{E}[(\sigma(a) - y)^2] = \mathbb{E}[(\sigma(a) - \mathbb{E}[y|a])^2] + \mathbb{E}[(y - \mathbb{E}[y|a])^2]$. Therefore, the supremum in (47) is achieved when $\sigma(a) = \mathbb{E}[y|a]$. Now (47) simplifies to

$$\mathsf{SR}_Z(r, \mathcal{D}) = \mathbb{E}_{(a,y)\sim\Gamma}[(a - \mathbb{E}[y|a])^2].$$

By the Cauchy-Schwarz inequality,

$$\mathsf{SR}_Z(r, \mathcal{D})\mathbb{E}[\varphi(a)^2] \geq \mathbb{E}[(a - \mathbb{E}[y|a])\varphi(a)]^2 = \mathbb{E}[(a - y)\varphi(a)]^2.$$ □

Proof of Theorem J.1. Consider an arbitrary distribution $\mathcal{D}$ on $[0,1]\times\{0,1\}$. Let Γ be the distribution of (a,y), where we first draw $(p,y)\sim\mathcal{D}$ and then draw $a\sim r(p)$. By Lemma J.2,

$$\mathsf{SR}_Z(r,\mathcal{D}) = \mathbb{E}_\Gamma[(a-\mathbb{E}[y|a])^2].$$

Let $\mathcal{D}$ be the distribution of (p,y) where $p=p_0\in[0,1-\varepsilon]$ is deterministic and $\mathbb{E}[y]=p_0+\varepsilon$. Clearly, we have $\mathsf{Cutoff}(\mathcal{D})=\varepsilon$. In this case, y and a are independent, so we have $\mathbb{E}[y|a]=p_0+\varepsilon$ and

$$\mathsf{SR}_Z(r,\mathcal{D}) = \mathbb{E}_{a\sim r(p_0)}[(a-(p_0+\varepsilon))^2] \geq \mathbb{E}_{a\sim r(p_0)}[(a-p_0)^2] - 2\varepsilon.$$

If $\mathbb{E}_{a\sim r(p_0)}[(a-p_0)^2] > 10^{-6}\varepsilon^{2/3}$, then we have $\mathsf{SR}_Z(r,\mathcal{D})\geq\Omega(\varepsilon^{2/3})$ and the theorem holds.

We can thus assume that $\mathbb{E}_{a\sim r(p)}[(a-p)^2]\leq 10^{-6}\varepsilon^{2/3}$ for every $p\in[0,1-\varepsilon]$. By Markov's inequality, for every $s>0$,

$$\Pr_{a\sim r(p)}[|a-p|\geq s] \leq \frac{10^{-6}\varepsilon^{2/3}}{s^2}. \tag{48}$$

We choose $m=\lfloor\varepsilon^{-1/3}\rfloor$ evenly spaced points $p_1,\dots,p_m\in[1/3,2/3]$, where each $p_i=1/3+i/(3m)$. We define the distribution $\mathcal{D}$ of (p,y) as follows. For every $i=1,\dots,m$, we set $\Pr[p=p_i]=3\varepsilon$. Thus the total probability mass on $\{p_1,\dots,p_m\}$ is $3m\varepsilon=\Theta(\varepsilon^{2/3})<1$. We place the remaining $1-3m\varepsilon$ probability mass on $p=0$. The label y is distributed as follows:

$$\mathbb{E}[y|p] = \begin{cases} 0, & \text{if } p=0;\\ p_i+(-1)^i/3, & \text{if } p=p_i \text{ for } i\in\{1,\dots,m\}.\end{cases}$$

It is easy to verify that $\mathsf{Cutoff}(\mathcal{D})=\varepsilon$ because the bias $(-1)^i/3$ in every two neighboring bins cancel out, so the bias of exactly one bin with weight 3ε contributes to $\mathsf{Cutoff}(\mathcal{D})$. It remains to show that $\mathsf{SR}_Z(r,\mathcal{D})=\Omega(\varepsilon^{2/3})$. We define

$$\varphi(a) := \sum_{i=1}^{m}(-1)^i\mathbb{I}[|a-p_i|<1/(6m)]\in\{-1,1\}.$$

Since $p_1,\dots,p_m\in[1/3,2/3]$,

$$\varphi(a)=0, \quad \text{for every } a\in[0,1/3-1/(6m)] \tag{49}$$

Now we bound the quantities related to φ in (46).

$$\begin{aligned}\mathbb{E}_\Gamma[\varphi(a)^2] &\leq \Pr[p\neq 0] + \Pr[p=0]\mathbb{E}_{a\sim r(0)}[\varphi(a^2)]\\ &\leq 3m\varepsilon + (1-3m\varepsilon)\Pr_{a\sim r(0)}[a\geq 1/3-1/(6m)] && \text{(by (49))}\\ &\leq O(\varepsilon^{2/3}). && \text{(by (48))}\end{aligned}$$

Also,

$$\mathbb{E}[(y-a)\varphi(a)|p=0] \geq -\Pr_{a\sim r(0)}[a\geq 1/3-1/(6m)] \geq -\varepsilon^{2/3}/100, \qquad \text{(by (49) and (48))}$$

and

$$\begin{aligned}\mathbb{E}[(y-a)\varphi(a)|p=p_i] &= \mathbb{E}[(p_i+(-1)^i/3-a)\varphi(a)|p=p_i]\\ &= \mathbb{E}_{a\sim r(p_i)}[(p_i+(-1)^i/3-a)\varphi(a)]\\ &\geq \Pr_{a\sim r(p_i)}[|a-p_i|<1/(6m)](1/3-1/(6m)) - \Pr_{a\sim r(p_i)}[|a-p_i|\geq 1/(6m)]\\ &\geq (9/10)(1/3-1/(6m))-1/10 && \text{(by (48))}\\ &\geq 1/10.\end{aligned}$$

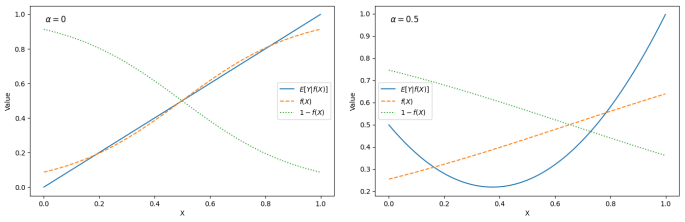
Combining the last two inequalities,

$$\mathbb{E}[(y-a)\varphi(a)] \geq -\varepsilon^{2/3}/100 + 3m\varepsilon/10 \geq \Omega(\varepsilon^{2/3}).$$

Plugging our bounds on $\mathbb{E}[\varphi(a)^2]$ and $\mathbb{E}[(y-a)\varphi(a)]$ into (46) proves that $\mathbb{E}[\mathsf{SR}_Z(r,\mathcal{D})]=\Omega(\varepsilon^{2/3})$. □

K Additional Experimental Results

To build intuition of the data distribution and the predictors used in our experiments, we plot $\mathbb{E}[Y|f(x)]$, $f(x)$ and $1-f(x)$ for different values of α in Figure 2.



(a) $\alpha = 0.0$ (b) $\alpha = 0.5$

(c) $\alpha = 0.8$ (d) $\alpha = 1$

Figure 2: Prediction curves for different values of α.

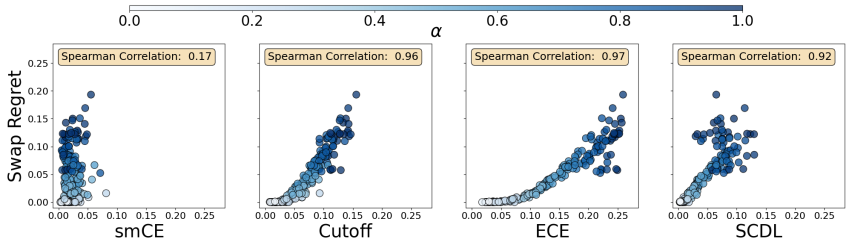


Figure 3: Plots of calibration error measures versus swap regret for predictor f. Spearman correlations are reported in each subplot.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We support the claims made in the abstract and the introduction by providing a technical overview in Section 2 and their full proofs in the appendix, and experimental results in Section 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Our contribution is primarily theoretical. The precise assumptions under which our results hold are clearly stated alongside each theorem. We discuss general limitations of SCDL in the introduction.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: We provide a technical overview of our contributions in Section 2 and their formal proofs in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setup and results are described in Section 3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code for our experiments is available at [https://github.com/konstantinabrk/scdl].

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe the experimental setting in full detail in Section 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For the testability experiment in Section 3 we report the standard deviation of the error measurements in Table 2. For the actionability experiment in Section 3, we report only the mean estimates of the calibration error measures without error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Given the simplicity of the experimental setup, we do not provide detailed information on the computer resources, as our experiments can be run on an average personal computer for a few minutes.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have read the code of ethics, and up to our best knowledge our paper abides by it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our contributions are theoretical and foundational in nature. In Section 1, we discuss how our framework enables downstream decision makers to trust and act on predictions, which has positive implications for reliable decision-making under uncertainty. Given the theoretical nature of our work, we do not foresee direct negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Most of the results of the paper were developed without the use of LLMs. However, as specified in the text, ChatGPT Pro was used to help discover the proof of Theorem J.1.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.