Toward Minimal-dimensional Convex Calibrated Surrogate Losses for Classification with Rejection

Yuzhou Cao[1] **Han Bao**[2,3,4] **Bo An**[1]

[1] College of Computing and Data Science, Nanyang Technological University
[2] The Institute of Statistical Mathematics
[3] Tohoku University
[4] RIKEN AIP
yuzhou002@e.ntu.edu.sg bao.han@ism.ac.jp
boan@ntu.edu.sg

Abstract

Classification with rejection (CwR) generalizes the standard classification task by providing an extra option of rejection with a cost c smaller than misclassification costs, and the design of its convex calibrated surrogate loss is our central topic. While typical surrogate losses for K-class classification operate on K-dimensional scoring functions, it has been proved that this dimensionality can be reduced to $\lceil \log_2 K \rceil$ for CwR when $c \in (0, 0.5]$, rendering more efficient computation. This logarithmic dependence is commonly conjectured to be the lowest possible among all convex calibrated surrogate losses for CwR. Through the lens of property elicitation, we explore even lower-dimensional structures for CwR, by constructing a 2-dimensional convex calibrated surrogate loss for $c \in (0, 0.5)$. This loss dimension is provably optimal among all convex calibrated surrogate losses. The boundary cost $c = 0.5$ is more nuanced—we construct a 4-dimensional convex calibrated loss, improving over the $\lceil \log_2 K \rceil$-dimension when $K > 16$. Notably, this breaks the optimal logarithmic dimensionality attainable by the *embedding framework* [Finocchiaro et al., 2024], a widely used loss design principle for polyhedral losses, negatively resolving their conjecture on the dimension optimality.

1 Introduction

In risk-critical classification scenarios, misclassification can be fatal. To address this issue, many studies have been devoted to classification with rejection (CwR), which allows a classifier to abstain from making high-risk decisions so as to avoid critical misclassification. Such an augmented classifier is commonly evaluated by the zero-one-c loss, which assigns a loss of 1 to misclassification as in standard classification and a fixed cost $c \in (0, 1)$ to rejection. Direct optimization of this target loss is intractable due to its discrete and discontinuous nature. A standard remedy is to employ *convex calibrated surrogate losses*, which enable tractable optimization and consistency, in the sense that minimizing the surrogate risk leads to minimizing the target risk.

Beyond convexity and calibration, another important factor for efficient optimization is the *dimensionality* of loss functions, which is the dimension of prediction scores evaluated by loss functions. A large loss dimension that is often caused by the combinatorial nature of discrete prediction tasks (e.g., extreme classification) can make optimization intractable. For this reason, a number of previous works have focused on lowering the dimensionality of convex calibrated surrogate losses, as well as its theoretical optimality [Finocchiaro et al., 2020, Ramaswamy et al., 2018, Finocchiaro et al., 2022]. The lowest possible dimensionality across all convex calibrated surrogate losses for a given task is called the *convex calibration dimension* of that task.

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

The topic of lower-dimensional surrogate losses for CwR has attracted considerable attention in the community. In particular, the BEP (binary encoding predictions) loss is a convex calibrated loss for the zero-one-c loss when $c \in (0, 0.5]$, achieving $\lceil \log_2 K \rceil$-dimension [Ramaswamy et al., 2018]. This logarithmic dependence on the class number is a substantial improvement over the standard $\mathcal{O}(K)$-dimensional calibrated surrogate losses for CwR, especially when encountering large class numbers. Furthermore, the tightness of this dimensionality is supported by existing lower bounds [Finocchiaro et al., 2020, 2021] for $K \in [2, 4]$. These results naturally lead to a conjecture in the community that this dimensionality is unimprovable, i.e., the convex calibration dimension of CwR is $\lceil \log_2 K \rceil$ [Finocchiaro et al., 2020, Conjecture 23].

In this work, we prove that this dimensionality can be further reduced to its minimally possible dimension, from the viewpoint of property elicitation [Steinwart et al., 2014, Agarwal and Agarwal, 2015]. In Section 3, we introduce a 2-dimensional property based on median and error-quantile estimation, which indirectly elicits the optimal decision rule of CwR. We then show in Section 4 that a 2-dimensional convex polyhedral loss can elicit this property and is therefore calibrated. This construction reduces the dimensionality from $\lceil \log_2 K \rceil$ to the constant 2, which is minimally possible (Corollary 15); that is, the convex calibration dimension of CwR is 2 for $c \in (0, 0.5)$. For $c = 0.5$, we show in Section 5 that the 2-dimensional property must be generalized to a 4-dimensional property for elicitability, and further construct a 4-dimensional convex calibrated loss based on this generalization in Section 6, strictly improving the existing logarithmic dimensionality when $K > 16$. Finally, in Section 7, we show that this improvement is impossible for any embedding-based loss by proving a matching logarithmic lower bound for CwR when $c = 0.5$ (Theorem 23). This negatively resolves Finocchiaro et al. [2020, Conjecture 23]. That is, the embedding dimension of a discrete loss can be strictly larger than its convex calibration dimension. Figure 1 summarizes the above results and their comparison with BEP.

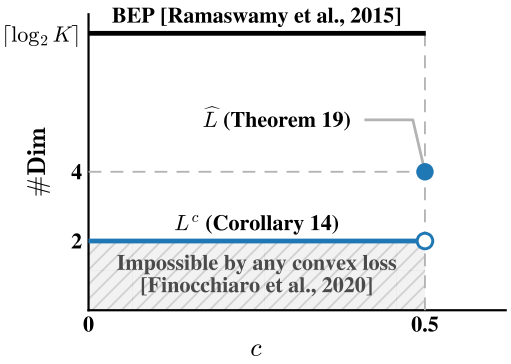


Figure 1: The dimensionality for convex calibrated CwR losses. Compared with the dimensionality of BEP, our results improve the dimensionality for both $c \in (0, 0.5)$ and $c = 0.5$, respectively in Corollary 14 and Theorem 19.

Related work. There is a rich literature on convex calibrated surrogate losses for CwR. Bartlett and Wegkamp [2008], Yuan and Wegkamp [2010], and Ni et al. [2019] propose hinge-like and class-probability-estimation-based losses, where rejection is determined by prediction scores and a threshold. In contrast, Cortes et al. [2016a, 2023] and Mao et al. [2024d] separate the rejection rule from prediction scores used for label prediction. Another line of work reduces CwR to ordinary classification problems [Charoenphakdee et al., 2021, Cao et al., 2022]. These losses typically have dimension $\mathcal{O}(K)$, motivating substantial interest in lower-dimensional constructions. Our key insight toward the loss-dimensionality reduction is to embed multi-class labels on the one-dimensional real line, shown in Section 3.1. This real-line encoding fundamentally differs from the standard simplex encoding of class labels used in the aforementioned CwR studies as well, yet is possible under CwR.

The first attempt to reduce the loss dimensionality for CwR dates back to Agarwal and Agarwal [2015], which is based on property elicitation of quantile vectors, with calibration guaranteed under a low-noise condition. The binary encoded predictions (BEP) loss [Ramaswamy et al., 2018] subsequently gives an assumption-free reduction to dimension $\lceil \log_2 K \rceil$ for $c \in (0, 0.5]$. Complementing these constructions, several theoretical works have established lower bounds on the dimension of convex calibrated losses for CwR [Finocchiaro et al., 2020, 2021], which match the logarithmic bound for certain class numbers. A more detailed review of related works is provided in Appendix A.

2 Preliminaries

Let $\Delta^K := \{\boldsymbol{\eta} \in \mathbb{R}^K_{\geq 0} : \|\boldsymbol{\eta}\|_1 = 1\}$ be the K-simplex, $[\![\cdot]\!]$ be the Iverson bracket, and $[K] := \{1, \ldots, K\}$. For a subset of the real line $\mathcal{S} \subseteq \mathbb{R}$ and $z \in \mathbb{R}$, we slightly abuse the notation by $\mathcal{S} \geq z$ whenever $s \geq z$ for all $s \in \mathcal{S}$. The p-norm is denoted by $\|\cdot\|_p$. For a random variable Z on $\mathbb{R}$, define its

$\alpha-$quantile by $Q_\alpha(Z) \coloneqq \{z \in \mathbb{R} : \Pr(Z < z) \leq \alpha,\ \Pr(Z \leq z) \geq \alpha\}$, and $\mathrm{Med}(Z) \coloneqq Q_{0.5}(Z)$ its median. We use Argmin/Argmax to denote the set of all minimizers/maximizers, and use argmin/argmax to denote the unique minimizer/maximizer, when these sets are singletons. Let $(x)_+ \coloneqq \max\{x, 0\}$, $(x)_- \coloneqq \max\{-x, 0\}$, $\rho_\alpha(\theta, y) \coloneqq \frac{\alpha}{1-\alpha}(y-\theta)_+ + (y-\theta)_-$ the pinball loss, and $\mathrm{CVaR}_\alpha(Z) \coloneqq \min_{r\in\mathbb{R}}[r + \frac{1}{1-\alpha}\mathbb{E}(Z-r)_+]$ the conditional value-at-risk (CVaR) of Z at α.

2.1 Classification with Rejection (CwR) and Calibrated Surrogate Losses

Let $\mathcal{Y} \coloneqq [K]$ be the label space, $\mathcal{X}$ the feature space, and $\boldsymbol{\eta}(\boldsymbol{x}) \in \Delta^K$ the distribution on $\mathcal{Y}$ with $\eta_y(\boldsymbol{x}) = \Pr(Y = y | X = \boldsymbol{x})$. In CwR, the prediction space is the augmented label space $\widehat{\mathcal{Y}} \coloneqq \mathcal{Y} \cup \{®\}$, where ® is the rejection option. The goal is to train an augmented classifier $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ that can abstain from making high-risk decisions, which is evaluated by the zero-one-c loss.

Definition 1 (Zero-one-c Loss)**.** *For a cost $c \in (0, 1)$, the* zero-one-c loss $\ell_{01c} : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ *is*

$$\ell_{01c}(t, y) = \begin{cases} [\![t \neq y]\!], & t \neq ®, \\ c, & \textit{otherwise}. \end{cases} \tag{1}$$

The loss ℓ_{01c} extends the *classification error (zero-one loss)* $\ell_{01}(t, y) \coloneqq [\![t \neq y]\!]$ for $t, y \in \mathcal{Y}$, by allowing the rejection option ® with an extra yet cheap cost c. The goal of CwR is to minimize the expected ℓ_{01c}, i.e., $\mathbb{E}_{X,Y}[\ell_{01c}(f(X), Y)] = \mathbb{E}_X[R_{\ell_{01c}}(f(X), \boldsymbol{\eta}(X))]$, and we call $R_{\ell_{01c}}(t, \boldsymbol{\eta}) \coloneqq \mathbb{E}_{Y\sim\boldsymbol{\eta}}[\ell_{01c}(t, Y)]$ 'the *zero-one-c risk*. The intuition of CwR can be fully characterized by the minimizer of this risk, which is called Chow's rule [Chow, 1970].

Lemma 2 (Chow's Rule)**.** *For any $\boldsymbol{\eta} \in \Delta^K$, the minimizers of $R_{\ell_{01c}}(t, \boldsymbol{\eta})$ are given as follows:*

$$\gamma_{01c}(\boldsymbol{\eta}) \coloneqq \mathrm{Argmin}_{t\in\widehat{\mathcal{Y}}} R_{\ell_{01c}}(t, \boldsymbol{\eta}) = \begin{cases} \{®\}, & \max_{y\in\mathcal{Y}} \eta_y < 1 - c, \\ \{®\} \cup \mathrm{Argmax}_{y\in\mathcal{Y}} \eta_y, & \max_{y\in\mathcal{Y}} \eta_y = 1 - c, \\ \mathrm{Argmax}\limits_{y\in\mathcal{Y}} \eta_y, & \textit{otherwise}. \end{cases} \tag{2}$$

This rule delineates that rejection is optimal if the maximal confidence falls below the threshold $1 - c$.

Given the discrete nature of ℓ_{01c}, it is often replaced with a convex surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$, which takes a real-vectored score $\boldsymbol{\theta} \in \mathbb{R}^d$ as a prediction instead of directly operating on the discrete prediction space $\widehat{\mathcal{Y}}$. Then a *prediction link function* $\varphi : \mathbb{R}^d \to \widehat{\mathcal{Y}}$ is further applied to connect the continuous score space with the discrete prediction space. For the pair of convex surrogate loss and prediction link, it is minimally necessary to guarantee that they are *calibrated* w.r.t. ℓ_{01c}.

Definition 3 (Calibration)**.** *A surrogate-prediction link pair (L, φ) is* calibrated *w.r.t. ℓ_{01c} if*

$$\inf_{\boldsymbol{\theta}\in\mathrm{dom}(L):\varphi(\boldsymbol{\theta})\notin\gamma_{01c}(\boldsymbol{\eta})} R_L(\boldsymbol{\theta}, \boldsymbol{\eta}) > \inf_{\boldsymbol{\theta}\in\mathrm{dom}(L)} R_L(\boldsymbol{\theta}, \boldsymbol{\eta}),\ \forall \boldsymbol{\eta} \in \Delta^K,$$

where $R_L(\boldsymbol{\theta}, \boldsymbol{\eta}) \coloneqq \mathbb{E}_{Y\sim\boldsymbol{\eta}}[L(\boldsymbol{\theta}, Y)]$ is the surrogate risk w.r.t. L.

In a nutshell, calibration (or consistency [Ramaswamy and Agarwal, 2016, Theorem 3]) requires that the surrogate risk is minimized if and only if we can recover Chow's rule from the transformed score $\varphi(\boldsymbol{\theta})$, which establishes the justification for the minimization of the zero-one-c risk.

A further studied target for convex calibrated surrogate losses is their minimally possible dimension [Ramaswamy and Agarwal, 2016, Finocchiaro et al., 2021], which is called the *convex calibration dimension*.

Definition 4 (Convex Calibration Dimension)**.** *The convex calibration dimension of ℓ_{01c} is the minimum dimension d for which there exists a convex set $\mathcal{C} \subseteq \mathbb{R}^d$, a convex surrogate loss $L : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}$, and a link $\varphi : \mathcal{C} \to \widehat{Y}$ such that (L, φ) is calibrated w.r.t. ℓ_{01c}.*

This concept was formally introduced in Ramaswamy and Agarwal [2016] for discrete prediction tasks, which is further extended to continuous-valued prediction tasks and the topic of property elicitation [Finocchiaro et al., 2021] that focuses on the unconstrained case of $\mathcal{C} = \mathbb{R}^d$.

2.2 Property Elicitation

Besides the calibration w.r.t. a target loss, convex loss functions have also been widely used for *property elicitation*, which aims to obtain a mapping defined on Δ^K by loss minimization.

Definition 5 (Property Elicitation)**.** *A property is a set-valued function $\Gamma : \Delta^K \rightrightarrows \mathcal{R}$, where $\mathcal{R}$ is a set of desired predictions or reports.*[1] *A loss $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$* (directly) elicits *a property Γ if we have*

$$\Gamma(\boldsymbol{\eta}) = \operatorname*{Argmin}_{\boldsymbol{\theta} \in \mathcal{R}} R_L(\boldsymbol{\theta}, \boldsymbol{\eta}), \quad \forall \boldsymbol{\eta} \in \Delta^K.$$

If a loss L elicits a property, the property is unique and we write the property by $\mathrm{prop}[L]$.

Chow's rule (2) can be regarded as the property elicited by the loss ℓ_{01c} with $\mathcal{R} = \hat{\mathcal{Y}}$. Another notion close to the context of calibration is indirect elicitation.

Definition 6 (Indirect Elicitation)**.** *For a loss function $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ and a link function $\varphi : \mathbb{R}^d \to \mathcal{R}$, the loss-link pair (L, φ)* indirectly elicits *a property $\gamma : \Delta^K \rightrightarrows \mathcal{R}$ if for any $\boldsymbol{\eta} \in \Delta^K$, we have $\varphi(\boldsymbol{\theta}) \in \gamma(\boldsymbol{\eta})$ for each $\boldsymbol{\theta} \in \mathrm{prop}[L](\boldsymbol{\eta})$.*

This concept is known to be a weaker notion compared with calibration: if γ is the property elicited by a loss ℓ, then the existence of an indirectly eliciting loss-link pair is necessary for the calibration of this pair w.r.t. ℓ [Finocchiaro et al., 2024]. Specifically, it is sufficient for calibration when L is polyhedral or convex differentiable with $d = 1$ [Khurana et al., 2025].

3 Lower-dimensional Structure of CwR through Property Elicitation

In this section, we turn to the possibility that CwR admits an even lower-dimensional intrinsic structure. We characterize this structure through the connection between Chow's rule and a 2-dimensional property. The proofs for this section are provided in Appendix B.

3.1 Chow's Rule through the Lens of Median and Quantile Elicitation

Prior loss-design approaches for CwR face a fundamental obstacle in constructing lower-dimensional surrogate losses: they often require eliciting the mode, $\mathrm{Argmax}_{y \in \mathcal{Y}} \eta_y$,[2] which is known to require convex losses of dimension at least $K - 1$ [Ramaswamy and Agarwal, 2016]. The binary encoded predictions (BEP) surrogate loss avoids this direct elicitation by reducing the problem to a collection of binary classification tasks induced by a $\lceil \log_2 K \rceil$-bit binary encoding [Ramaswamy et al., 2018]. However, its dimensionality still grows with the number of classes K.

In this section, we show that both the mode required for CwR and a rejection rule consistent with Chow's rule can be recovered from a 2-dimensional property. After embedding the categorical random variable $Y \in [K] \subset \mathbb{R}$ into the real line $\mathbb{R}$ via the canonical injection, the discrete mode of Y can be alternatively elicited via a continuous 1-dimensional property, namely, the **median** of Y. This median–mode equivalence holds under some conditions, shown as follows.

Lemma 7 (Median as Mode)**.** *Fix any $\boldsymbol{\eta} \in \Delta^K$ such that the* strong dominant-label condition $\max_{y \in \mathcal{Y}} \eta_y > 0.5$ *holds. For a random variable Y supported on the real line, suppose that $\Pr(Y = y) = \eta_y$ for $y \in [K]$ and $\Pr(Y \notin [K]) = 0$. Then, we have $\mathrm{Med}(Y) = \{\mathrm{argmax}_{y \in \mathcal{Y}} \eta_y\}$.*

This conclusion follows from the fact that both $\Pr(Y \geq y^*)$ and $\Pr(Y \leq y^*)$ are at least $\Pr(Y = y^*) = \eta_{y^*} > 0.5$, where $y^* := \mathrm{argmax}_{y \in \mathcal{Y}} \eta_y$. This median–mode equivalence is important for constructing lower-dimensional surrogate losses. Since the median is known to be elicitable by 1-dimensional regression losses, such as the quantile (pinball) loss [Christmann and Steinwart, 2007], Lemma 7 shows that the mode can be indirectly elicited through the median (1-dimensional property) whenever the strong dominant-label condition holds.

This strong dominant-label condition partially aligns with the interest of CwR when $c \in (0, 0.5)$. Indeed, Chow's rule (2) predicts $\mathrm{argmax}_{y \in \mathcal{Y}} \eta_y$ whenever $\max_{y \in \mathcal{Y}} \eta_y > 1 - c \; (> 0.5)$. However,

[1] A *report* is a synonym of a prediction, as well as a (prediction) score, throughout the paper. In the viewpoint of property elicitation, a report lives in the codomain of a property Γ.

[2] The mode is directly elicited by the zero-one loss ℓ_{01}, which is the target loss of multi-class classification.

this 1-dimensional property is still insufficient to fully recover Chow's rule. In particular, the median alone does not reveal whether $\max_{y\in\mathcal{Y}} \eta_y \geq 1-c$, and thus cannot determine whether rejection is required.

We view Chow's rejection rule from a different angle through the lens of the *quantile*. Noting that the **classification error** ℓ_{01} elicits the mode, Chow's rejection rule $\max_y \eta_y \leq 1-c$ is satisfied when the minimal expected classification error $\min_{t\in\mathcal{Y}} R_{\ell_{01}}(t, \boldsymbol{\eta}) = 1 - \max_{y\in\mathcal{Y}} \eta_y$ exceeds the threshold c, where $R_{\ell_{01}}(t, \boldsymbol{\eta}) \coloneqq \mathbb{E}_{Y\sim\boldsymbol{\eta}}[\ell_{01}(t, Y)] = 1 - \eta_t$. As the classification error is discrete and nonconvex, we attempt to replace it with the absolute deviation loss $|\theta - Y|$ (for a 1-dimensional continuous score $\theta \in \mathbb{R}$), which elicits $\mathrm{Med}(Y)$. It turns out that the absolute deviation can also be related to the rejection rule.

Lemma 8 (Quantile of Error as Rejection Rule)**.** *For any $c \in (0, 0.5]$, $\boldsymbol{\eta} \in \Delta^K$, and $Y \sim \boldsymbol{\eta}$:*

- *If $\max_{y\in\mathcal{Y}} \eta_y > 1-c$, then $Q_{1-c}(|\theta - Y|) = \{0\}$ for $\theta \in \mathrm{Med}(Y) = \{\mathrm{argmax}_{y\in\mathcal{Y}} \eta_y\}$.*
- *If $\max_{y\in\mathcal{Y}} \eta_y < 1-c$, then $Q_{1-c}(|\theta - Y|) \geq 0.5$ for any $\theta \in \mathbb{R}$.*
- *If $\max_{y\in\mathcal{Y}} \eta_y = 1-c$, then $Q_{1-c}(|\theta - Y|) \geq \min\{0.5, \min_{y^*\in \mathrm{Argmax}_y \eta_y} |\theta - y^*|\}$ for any $\theta \in \mathbb{R}$.*

The first two cases of Lemma 8 indicate that given the median $\theta \in \mathrm{Med}(Y)$ w.r.t. $\boldsymbol{\eta}$, we should reject whenever $Q_{1-c}(|\theta - Y|)$ is larger than 0.5 (and vice versa). Interestingly, according to the second case, θ is not required to be the median on the rejected side $\max_y \eta_y < 1-c$, which closely matches Chow's rule. When rejection is optimal, the label prediction can be arbitrary, and so is the score θ.

For each $\boldsymbol{\eta} \in \Delta^K$, the minimal necessary information required by Lemmas 7 and 8 can be formulated as the following property:

$$\widetilde{\Gamma}(\boldsymbol{\eta}) = \begin{cases} \left\{[\theta, r]^\top \in \mathbb{R}^2 : \ \theta \in \mathrm{Med}(Y), r \in Q_{1-c}(|\theta - Y|)\right\}, & \max_{y\in\mathcal{Y}} \eta_y > 1-c, \\ \left\{[\theta, r]^\top \in \mathbb{R}^2 : \ \theta \in \mathbb{R}, r \in Q_{1-c}(|\theta - Y|)\right\}, & \max_{y\in\mathcal{Y}} \eta_y \leq 1-c. \end{cases} \tag{3}$$

Thus, $\widetilde{\Gamma}$ is designed so that its second coordinate detects the rejection region, while its first coordinate provides a label prediction whenever the input is not rejected by r. The following lemma shows that we can recover Chow's rule from the property $\widetilde{\Gamma}$ via a deterministic link.

Lemma 9 (Recovery of Chow's Rule from $\widetilde{\Gamma}$)**.** *Define $\varphi^\tau : \mathbb{R}^2 \to \widehat{\mathcal{Y}}$ by*

$$\varphi^\tau([\theta, r]^\top) = \begin{cases} \text{®}, & r > \tau, \\ \widetilde{\mathrm{argmin}}_{y\in\mathcal{Y}} |\theta - y|, & \textit{otherwise}, \end{cases} \tag{4}$$

where $\widetilde{\mathrm{argmin}}_{y\in\mathcal{Y}}$ is an arbitrary deterministic selection from the set $\mathrm{Argmin}_{y\in\mathcal{Y}} |\theta - y|$. When $\tau \in [0, 0.5)$ and $c \in (0, 0.5]$, we can recover Chow's rule from $\varphi^\tau \circ \widetilde{\Gamma}$, i.e., for any $\boldsymbol{\eta} \in \Delta^K$,

$$\varphi^\tau([\theta, r]^\top) \in \gamma_{01c}(\boldsymbol{\eta}), \ \forall [\theta, r]^\top \in \widetilde{\Gamma}(\boldsymbol{\eta}).$$

Lemma 9 paves a clear path toward the design of a lower-dimensional convex calibrated surrogate loss for CwR: if we can design a convex loss L such that elicits $\widetilde{\Gamma}$ or its non-empty subset Γ (in the sense that $\Gamma(\boldsymbol{\eta}) \subseteq \widetilde{\Gamma}(\boldsymbol{\eta})$) and is non-empty for any $\boldsymbol{\eta} \in \Delta^K$, then this convex loss L is naturally 2-dimensional and further **indirectly elicits** Chow's rule! Therefore, such a 2-dimensional loss L would be immediately calibrated with ℓ_{01c} because indirect elicitation is a necessary condition (and can be made sufficient with additional conditions) for calibration [Finocchiaro et al., 2024].

4 Minimal-dimensional Convex Calibrated Surrogate Losses for CwR when $c < 0.5$

In Section 3, we proved that we can recover Chow's rule from the 2-dimensional property $\widetilde{\Gamma}$ defined in (3), which gives a recipe of a 2-dimensional convex calibrated surrogate loss for CwR. However, its compositional structure makes the elicitation of this property not straightforward. In this section, we consider elicitation of a subset of $\widetilde{\Gamma}$ instead, which leads to a 2-dimensional convex calibrated loss. The proofs for this section are provided in Appendix C.

4.1 Surrogate Loss Construction and Property Analysis

Conic combination of mean absolute and pinball losses. Consider eliciting $\widetilde{\Gamma}$ in (3), which consists of $\mathrm{Med}(Y)$ and the quantile of the absolute error $|\theta - Y|$. Typically, a median and quantile are estimated by minimizing the mean absolute and pinball losses, respectively. To elicit $\widetilde{\Gamma}$, we need to elicit them jointly, for which we consider the following conic combination of the two losses:

$$L^c([\theta, r]^\top, y) = \underbrace{\left[r + \frac{1}{c}\max(|\theta - y| - r, 0)\right]}_{\text{for error quantile elicitation ("CVaR objective")}} + \lambda \underbrace{|\theta - y|}_{\text{for median elicitation}}, \text{ for a given } \lambda > 0. \quad (5)$$

This loss is convex and polyhedral in $[\theta, r]^\top$. The second term $|\theta - y|$ is the standard mean absolute loss used to elicit $\mathrm{Med}(Y)$. In contrast, the first term is intimately related to the pinball loss—in fact, its partial minimum of expectation $\min_r \left(r + \frac{1}{c}\mathbb{E}\max(|\theta - Y| - r, 0)\right)$ is the $(1-c)$-CVaR objective, and both CVaR and the expected pinball loss are minimized at the quantile $r \in Q_{1-c}(|\theta - Y|)$, which is the fact known as a part of "the fundamental risk quadrangle" [Rockafellar and Uryasev, 2013].

Median–quantile joint elicitation. While the mean absolute and pinball losses *alone* elicit median and quantile, respectively, it is not straightforward to see whether the conic combination loss elicits the two properties *jointly*. Specifically, the elicited median $\theta^* \in \mathrm{Med}(Y)$ via the mean absolute loss can cascade down to the CVaR objective in (5), which complicates the optimality analysis jointly w.r.t. $[\theta, r]^\top$. While joint optimality is intricate in general, our conic surrogate loss L^c has a better optimality structure thanks to the interaction between mean and quantile. As we see subsequently, we can guarantee the quantile $r \in Q_{1-c}(|\theta - Y|)$ is optimal, successfully eliciting $\widetilde{\Gamma}$, as long as the conic combination coefficient λ is sufficiently large.

Theorem 10. *For any $\boldsymbol{\eta} \in \Delta^K$ and $c \in (0, 0.5)$, if $\lambda > \frac{1}{1-2c}$, then $\mathrm{prop}[L^c](\boldsymbol{\eta}) \neq \emptyset$ and $\mathrm{prop}[L^c](\boldsymbol{\eta}) \subseteq \widetilde{\Gamma}(\boldsymbol{\eta})$ for any $\boldsymbol{\eta} \in \Delta^K$. Furthermore, (L^c, φ^τ) indirectly elicits γ_{01c} for $\tau \in (0, 0.5)$.*

Proof sketch. Consider the optimality of the surrogate risk $R_{L^c}([\theta, r]^\top, \boldsymbol{\eta}) = \mathbb{E}_{Y\sim\boldsymbol{\eta}}[L^c([\theta, r]^\top, Y)]$. We need to check if the optimal $[\theta, r]^\top$ is included in $\widetilde{\Gamma}(\boldsymbol{\eta})$. Let us focus on the optimality w.r.t. r solely. For any fixed θ, the optimal r satisfies $r \in Q_{1-c}(|\theta - Y|)$. Write the surrogate risk with this r by

$$\widetilde{R}_{L^c}(\theta, \boldsymbol{\eta}) \coloneqq \min_{r\in\mathbb{R}} R_{L^c}([\theta, r]^\top, \boldsymbol{\eta}) = \mathrm{CVaR}_{1-c}(|\theta - Y|) + \lambda\mathbb{E}_Y[|\theta - Y|]. \quad (6)$$

When $\max_y \eta_y > 1 - c$ and $\theta \in \mathrm{Med}(Y) = \{\mathrm{argmax}_y \eta_y\}$, the subdifferential of $\mathbb{E}_Y[|\theta - Y|]$ contains 0 in its interior, which implies that $\partial_\theta \mathbb{E}_Y[|\theta - Y|]$ also contains both positive and negative subgradients. Thus, regardless of the subgradient of CVaR, we can ensure $\mathrm{int}(\partial_\theta \widetilde{R}_{L^c}(\theta, \boldsymbol{\eta})) \ni 0$ by increasing λ to offset the subgradient of CVaR, and $\lambda > \frac{1}{1-2c}$ is the necessary condition to allow this offset. In this case, θ is the sharp and thus unique minimizer [Polyak, 1987, Lemma 5.3.3]. □

Remark 11 (Convexity Issues of the Quantile Estimation Loss)**.** *To elicit the quantile of $|\theta - Y|$, we adopt the CVaR-based loss function because this yields a convex objective and the subgradient analysis w.r.t. θ is sufficient for the optimality. The common pinball loss [Koenker and Bassett, 1978, Christmann and Steinwart, 2007, Steinwart et al., 2014] becomes nonconvex in our case because the random variable $|\theta - Y|$ contains the optimizable parameter θ. As a result, the partial minimum of the pinball loss w.r.t. r yields a nonconvex function of θ [Rockafellar and Uryasev, 2013].*

4.2 Calibration Analysis

Although Theorem 10 proved that (L^c, φ^τ) indirectly elicits Chow's rule, it does not yet establish the calibration of L^c, since indirect elicitation is only necessary for calibration [Finocchiaro et al., 2021, Proposition 1]. In this subsection, we establish calibration through the notion of *separated link*, which strengthens indirect elicitation from the geometric perspective and is crucial for the calibration analysis of polyhedral losses.

Definition 12 (Separated Link [Finocchiaro et al., 2024, Definition 16])**.** *For two properties $\Gamma : \Delta^K \rightrightarrows \mathbb{R}^d$ and $\gamma : \Delta^K \rightrightarrows \mathcal{R}$, a link function $\varphi : \mathbb{R}^d \to \mathcal{R}$ is said to be a* ϵ-separated link *w.r.t. Γ and γ if $\inf_{\boldsymbol{\theta}' \in \Gamma(\boldsymbol{\eta})} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty \geq \epsilon$ for any $\boldsymbol{\eta} \in \Delta^K$ and $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $\varphi(\boldsymbol{\theta}) \notin \gamma(\boldsymbol{\eta})$.*

This concept imposes a stronger geometric requirement than indirect elicitation. In fact, indirect elicitation only requires that the surrogate property $\Gamma(\boldsymbol{\eta})$ does not intersect the set of surrogate reports that are linked to "incorrect" target decisions, $\{\boldsymbol{\theta} \in \mathbb{R}^d : \varphi(\boldsymbol{\theta}) \notin \gamma(\boldsymbol{\eta})\}$, while a separated link further requires these two sets to be uniformly separated by a strictly positive distance.

For polyhedral losses, it is always possible to construct a separated link by extending a given link which indirectly elicits a discrete property [Finocchiaro et al., 2024, Construction 1]. In our work, we obtain such an extension by restricting the range of parameter τ in the link φ^τ in (4) as follows:

Lemma 13. *For any $c \in (0, 0.5)$, φ^τ is $\min(\tau, 0.5 - \tau)$-separated w.r.t. $\mathrm{prop}[L^c]$ and γ_{01c} if $\lambda > \frac{1}{1-2c}$ and $\tau \in (0, 0.5)$.*

Unlike the range $\tau \in [0, 0.5)$ in (4), the modified range prohibits the choice $\tau = 0$ when we ensure the separability. Intuitively, the rejection region given by the link (4) becomes $r > 0$ when $\tau = 0$. However, by Lemma 8 and the property (3), $r = 0$ is optimal when $\max_y \eta_y > 1 - c$. Thus, rejected reports can approach an optimal accepted report arbitrarily closely by keeping θ fixed and letting r approach 0, so the link is not separated when $\tau = 0$. Since L^c is polyhedral, Lemma 13 directly implies calibration [Finocchiaro et al., 2024, Theorem 17].

Corollary 14. *For any $c \in (0, 0.5)$ and $\tau \in (0, 0.5)$, (L^c, φ^τ) is calibrated w.r.t. ℓ_{01c} if $\lambda > \frac{1}{1-2c}$.*

Therefore, we successfully demonstrate the existence of a 2-dimensional calibrated surrogate loss L^c for CwR. Furthermore, according to Finocchiaro et al. [2020, Theorem 12, Proposition 20], there exists no 1-dimensional convex calibrated surrogate losses for multi-class CwR when $c < 1$, which indicates that L^c has achieved the minimally possible dimensionality.[3]

Corollary 15. *For $K \geq 3$, the convex calibration dimension of ℓ_{01c} when $c \in (0, 0.5)$ is 2.*

5 Failure of the Lower-dimensional Structure When $c = 0.5$

We proposed a 2-dimensional convex calibrated surrogate loss for CwR by eliciting subsets of the property $\widetilde{\Gamma}$ as an intermediate step. However, the calibration result remains unclear when $c = 0.5$. In particular, Corollary 14 requires $\lambda > \frac{1}{1-2c}$, a condition that is not meaningful when $c = 0.5$. This raises a natural question: is the lower-dimensional structure characterized by $\widetilde{\Gamma}$ in Section 3 still valid for loss design when $c = 0.5$? Unfortunately, the following conclusion gives a negative answer. Specifically, no subset of $\widetilde{\Gamma}$ can be elicited by a 2-dimensional convex loss in this case.

Theorem 16 (Impossibility of Convex Elicitation)**.** *When $c = 0.5$ and $K \geq 2$, there is no 2-dimensional convex loss L such that $\mathrm{prop}[L](\boldsymbol{\eta}) \subseteq \widetilde{\Gamma}(\boldsymbol{\eta})$ for any $\boldsymbol{\eta} \in \Delta^K$.*

The proof is provided in Appendix D. To better understand the obstruction specific to our 2-dimensional construction, recall from Lemma 7 that, under the strong dominant-label condition $\max_y \eta_y > 0.5$, the median of Y coincides with the mode. Chow's rule predicts a label when $\max_y \eta_y > 1 - c (> 0.5)$, and therefore the median estimation is sufficient for recovering the Bayes-optimal decision as long as $c < 0.5$. The case $c = 0.5$ is fundamentally different. In this case, Chow's rule predicts a label when $\max_y \eta_y \geq 0.5$. When the equality $\max_y \eta_y = 0.5$ is satisfied, the median is not guaranteed to recover the mode, thus median estimation may fail to recover Chow's rule.

6 Constant-dimensional Convex Calibrated Surrogates Losses for CwR with $c = 0.5$

Our construction of the 2-dimensional surrogate loss L^c in Section 3 proceeds in two steps: first, L^c is required to elicit a subset of $\widetilde{\Gamma}$; then, calibration is established through the separated link φ^τ. However, Theorem 16 shows that the first step fails when $c = 0.5$. Consequently, the 2-dimensional structure characterized by $\widetilde{\Gamma}$ does not directly inform the design of 2-dimensional convex calibrated

[3] While Finocchiaro et al. [2020, Theorem 12, Proposition 20] show the nonexistence of 1-dimensional surrogate losses by supposing that the surrogate loss domain is unconstrained such that $\mathcal{C} = \mathbb{R}^d$, we show in Appendix G that the constrained $\mathcal{C}$ in Definition 4 is equivalent to the unconstrained case when $d = 1$. Therefore, Finocchiaro et al.'s impossibility is applicable and directly leads to Corollary 14.

losses for CwR in the case $c = 0.5$. Does this observation rule out any lower-dimensional structures for CwR at $c = 0.5$? In this section, we explore lower-dimensional convex calibrated losses than the BEP loss.

6.1 A 4-dimensional Structure of CwR with $c = 0.5$

In Section 5, we saw that it is challenging to derive a 2-dimensional convex calibrated loss when $c = 0.5$ because the median–mode equivalence breaks down. This challenge leads to the impossibility of eliciting Chow's rule if convex losses have dimensionality no more than 2, by combining Theorem 16 and Finocchiaro et al. [2020, Proposition 20].

Here, we seek a slightly higher- yet constant-dimensional space to overcome the technical challenges when $c = 0.5$. The following lemma is a key observation.

Lemma 17. *Fix any $\boldsymbol{\eta} \in \Delta^K$ such that the* weak dominant-label condition $\max_y \eta_y \geq 0.5$ *holds, consider a random variable $Y \sim \eta$ and fix $\delta \in (0, 0.25)$. Then, for any $y^* \in \operatorname{Argmax}_y \eta_y$, we have $Q_{0.5+\delta}(Y) = \{y^*\}$ or $Q_{0.5-\delta}(Y) = \{y^*\}$.*

This lemma shows that eliciting the $(0.5 \pm \delta)$-quantiles of Y is sufficient to recover an exact mode of Y when $\max_{y \in \mathcal{Y}} \eta_y \geq 0.5$. This *weak* dominant-label condition relaxes the *strong* dominant-label condition $\max_{y \in \mathcal{Y}} \eta_y > 0.5$ in Lemma 7. Therefore, we can indirectly elicit the mode via both $Q_{0.5 \pm \delta}(Y)$, which is less dimension-efficient compared with eliciting the median only, but with relaxing the strong dominant-label condition. The weak dominant-label condition in Lemma 17 is the crux to circumvent the challenges arising from $c = 0.5$.

To recover the rejection rule in Chow's rule, a useful observation is that the characterization in Lemma 8 remains valid when $c = 0.5$. In particular, for an appropriate report θ, the quantity $Q_{1-0.5}(|\theta - Y|) = \operatorname{Med}(|\theta - Y|)$ can still serve as a rejection rule, as in Section 3.

These observations suggest that, instead of estimating $\operatorname{Med}(Y)$ directly, one should elicit $Q_{0.5-\delta}(Y)$ and $Q_{0.5+\delta}(Y)$, together with the medians of their corresponding absolute deviations. That is, it is sufficient to elicit the 4-dimensional property to recover Chow's rule when $c = 0.5$. In the next subsection, we construct a 4-dimensional convex loss based on this structure and show it yields calibration w.r.t. ℓ_{01c} when $c = 0.5$.

6.2 Loss Formulation and Calibration

Inspired by the connection between $Q_{0.5 \pm \delta}(Y)$ and $\operatorname{Med}(|\theta - Y|)$, we construct a 4-dimensional loss that aims to estimate these four statistics as formulated below:

$$\widehat{L}([\theta_1, r_1, \theta_2, r_2]^\top, y) = L^+([\theta_1, r_1]^\top, y) + L^-([\theta_2, r_2]^\top, y), \tag{7}$$

where L^+ and L^- are conic combinations of quantile estimation losses with positive weights λ^+ and λ^-:

$$L^+([\theta, r]^\top, y) = \underbrace{r + 2(|y - \theta| - r)_+}_{\text{for error quantile elicitation}} + \lambda^+ \underbrace{\rho_{0.5+\delta}(\theta, y)}_{\text{for label quantile elicitation}}, \tag{8}$$

$$L^-([\theta, r]^\top, y) = \underbrace{r + 2(|y - \theta| - r)_+}_{\text{for error quantile elicitation}} + \lambda^- \underbrace{\rho_{0.5-\delta}(\theta, y)}_{\text{for label quantile elicitation}}. \tag{9}$$

Compared with the surrogate loss (5), we replace the mean absolute loss with the pinball losses in (8) and (9) to estimate the $(0.5 \pm \delta)$-quantiles of Y instead of its median. Meanwhile, the same CVaR-based losses are used to estimate the quantile of the absolute error. By minimizing the risk $R_{\widehat{L}}([\theta_1, r_1, \theta_2, r_2]^\top, \boldsymbol{\eta})$, we aim to reconstruct Chow's rule. The following theorem justifies this construction.

Theorem 18. *Set $\delta \in (0, 0.25)$, $\lambda^+ > \frac{0.5-\delta}{\min(\delta, 0.25-\delta)}$, and $\lambda^- > \frac{0.5+\delta}{\min(\delta, 0.25-\delta)}$. Then $\operatorname{prop}[\widehat{L}](\boldsymbol{\eta})$ is non-empty for any $\boldsymbol{\eta} \in \Delta^K$. Furthermore, for any $[\theta_1^*, r_1^*, \theta_2^*, r_2^*]^\top \in \operatorname{prop}[\widehat{L}](\boldsymbol{\eta})$:*

- *If $\max_y \eta_y < 0.5$: $r_i^* \geq 0.5$ for $i \in \{1, 2\}$.*
- *If $\max_y \eta_y > 0.5$: $(\theta_1^*, r_1^*) = (y^*, 0)$ or $(\theta_2^*, r_2^*) = (y^*, 0)$, where $y^* \coloneqq \operatorname{argmax}_{y \in \mathcal{Y}} \eta_y$.*

- *If* $\max_y \eta_y = 0.5$*:* $r_i^* \geq \min(0.5, \min_{y' \in \operatorname{Argmax}_y \eta_y} |\theta_i^* - y'|)$ *for* $i \in \{1, 2\}$.

If $\boldsymbol{\eta}$ lies in the rejection region $\max_y \eta_y \leq 0.5$, both r_1^* and r_2^* are at least 0.5; if $\boldsymbol{\eta}$ lies in the acceptance region, (i.e., $\max_y \eta_y > 0.5$), then at least one of (r_1^*, r_2^*) is equal to zero. Thus, the acceptance and rejection regions are separated with a strictly positive margin in terms of (r_1^*, r_2^*). Moreover, r_i^* also indicates which θ_i^* should be used for prediction: θ_i^* with the smaller r_i is the reliable candidate for recovering the accepted label. With these observations, we can recover Chow's rule from $\operatorname{prop}[\widehat{L}]$.

The following theorem provides a separated link associated with $\operatorname{prop}[\widehat{L}]$.

Theorem 19. *For* $c = 0.5$*,* $\delta \in (0, 0.25)$*,* $\lambda^+ > \frac{0.5-\delta}{\min(\delta, 0.25-\delta)}$*,* $\lambda^- > \frac{0.5+\delta}{\min(\delta, 0.25-\delta)}$*, and* $\tau \in (0, 0.5)$*, the following link* $\hat{\varphi}^\tau$ *is* $\min(\tau, \frac{0.5-\tau}{2})$*-separated w.r.t.* $\operatorname{prop}[\widehat{L}]$ *and* γ_{01c}*, and* $(\widehat{L}, \hat{\varphi}^\tau)$ *is calibrated w.r.t.* ℓ_{01c} *for* $c = 0.5$*.*

$$\hat{\varphi}^\tau([\theta_1, r_1, \theta_2, r_2]^\top) \coloneqq \begin{cases} \widetilde{\operatorname{argmin}}_y |\theta_1 - y|, & r_1 \leq \tau, \\ \widetilde{\operatorname{argmin}}_y |\theta_2 - y|, & r_1 > \tau,\ r_2 \leq \tau, \\ \text{®}, & \textit{otherwise}. \end{cases} \tag{10}$$

Compared with the BEP loss, the dimensionality of $\widehat{L}$ is no larger than $\lceil \log_2 K \rceil$ when $K > 8$ and is strictly smaller when $K > 16$, yielding a non-trivial improvement. Thus, our results also disprove the conjecture that $\lceil \log_2 K \rceil$ is the convex calibration dimension of ℓ_{01c} [Finocchiaro et al., 2020].

Remark 20. *While the losses* (5) *and* (7) *are designed primarily to investigate the CCdim of CwR, their calibration and polyhedral structure also guarantee linear surrogate regret bounds [Finocchiaro et al., 2024, Theorem 19]. These bounds allow statistical guarantees on surrogate regret/excess risk to transfer linearly to the target regret.*

7 Embedding Dimension Is Larger Than Convex Calibration Dimension

In the topic of dimension reduction of convex losses, polyhedral (i.e., convex and piecewise affine) losses play a central role. Moreover, the calibration of polyhedral losses is relevant to a geometrically equivalent concept called *embedding*, which is defined below.

Definition 21 (Embedding [Finocchiaro et al., 2019])**.** *A loss* $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ *embeds* $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ *if there exists an embedding* $\boldsymbol{\psi} : \mathcal{R} \to \mathbb{R}^d$ *such that (i) for all* $r \in \mathcal{R}$ *and* $y \in \mathcal{Y}$ *we have* $L(\boldsymbol{\psi}(r), y) = \ell(r, y)$*, and (ii) for all* $\boldsymbol{\eta} \in \Delta^K$ *and* $r \in \mathcal{R}$ *we have* $r \in \operatorname{prop}[\ell](\boldsymbol{\eta}) \Leftrightarrow \boldsymbol{\psi}(r) \in \operatorname{prop}[L](\boldsymbol{\eta})$*.*

This concept allows an embedding to map a discrete report $r \in \mathcal{R}$ into the d-dimensional Euclidean space, where d can be chosen as small as possible as long as the structure of ℓ is preserved by the corresponding loss L in this space. The smallest possible d is called the embedding dimension.

Definition 22 (Embedding Dimension [Finocchiaro et al., 2019])**.** *The embedding dimension of* ℓ *is the minimum dimension* d *such that there exists a polyhedral convex* $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ *that embeds* ℓ*.*

It is shown in Finocchiaro et al. [2020] that the embedding dimension upper-bounds the convex calibration dimension, and it is also pointed out that the embedding dimension is substantially simpler to analyze, thereby motivating the study of the embedding dimension as a proxy for the convex calibration dimension. Further motivated by the success of embedding-based losses in loss dimension reduction [Ramaswamy et al., 2018, Finocchiaro et al., 2019, 2022], Finocchiaro et al. [2020, Conjecture 23] conjectured that the embedding dimension of any discrete loss coincides with its convex calibration dimension. If true, this conjecture would considerably simplify the analysis of the convex calibration dimension.

However, we show that this conjecture fails by establishing the exact embedding dimension of ℓ_{01c} for $c = 0.5$. This result implies that the BEP loss achieves the lowest possible dimensionality among convex surrogate losses that embed ℓ_{01c}.

Theorem 23. *For a convex surrogate loss* $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ *that embeds* ℓ_{01c} *for* $c = 0.5$*, we have* $d \geq \lceil \log_2 K \rceil$*, and thus the embedding dimension for* ℓ_{01c} *when* $c = 0.5$ *is exactly* $\lceil \log_2 K \rceil$*.*

It follows from the embedding dimension analysis and Theorem 19 that the embedding dimension of ℓ_{01c} when $c = 0.5$ is strictly larger than its convex calibration dimension with a large enough K. This separation disproves Finocchiaro et al. [2020, Conjecture 23].

Theorem 24. *When $c = 0.5$ and $K > 16$, the embedding dimension of ℓ_{01c} is strictly larger than its convex calibration dimension.*

Remark 25. *By Theorem 23, our proposed 4-dimensional surrogate loss $\widehat{L}$ does **not** embed ℓ_{01c}. This explains how it bypasses the $\lceil \log_2 K \rceil$ lower bound imposed on embedding-based constructions. Although the construction of L^c and $\widehat{L}$ does not rely on the embedding framework, their success in reducing surrogate losses' dimension still demonstrates the strength of the polyhedral loss framework. Notably, the possibility of constructing calibrated polyhedral losses without direct embedding has been discussed by Finocchiaro et al. [2024, Sec. 7], and further developing this perspective is a promising direction for future work.*

8 Conclusions

In this paper, we develop lower-dimensional convex calibrated surrogate losses for CwR through the lens of median and quantile estimation. Specifically, for ℓ_{01c} when $c \in (0, 0.5)$, we give an explicit 2-dimensional convex calibrated surrogate loss, improving upon the existing $\lceil \log_2 K \rceil$-dimensional loss [Ramaswamy et al., 2018]. We further prove that this dimension is minimal among all convex calibrated surrogate losses, thereby establishing that the convex calibration dimension of ℓ_{01c} when $c \in (0, 0.5)$ is exactly 2. For $c = 0.5$, we propose a 4-dimensional convex calibrated surrogate loss, which improves upon the $\lceil \log_2 K \rceil$-dimensional loss when $K > 16$. We also show that such a reduction is impossible for embedding-based losses [Finocchiaro et al., 2024]: for ℓ_{01c} when $c = 0.5$, every loss in this family must have dimension at least $\lceil \log_2 K \rceil$. This lower bound further determines the exact embedding dimension of ℓ_{01c} in this case, and refutes the conjecture that embedding dimension always coincides with convex calibration dimension [Finocchiaro et al., 2020]. An important remaining question is whether $\widehat{L}$ achieves the convex calibration dimension of ℓ_{01c} when $c = 0.5$, which we leave as a promising direction for future work.

Acknowledgment

YC is supported by the Google PhD Fellowship program. HB is supported by JST PRESTO JPMJPR24K6.

References

Jacob D. Abernethy and Rafael Frongillo. A characterization of scoring rules for linear properties. In *Conference on Learning Theory*, page 27.1, 2012. Cited on page: 16

Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *Conference on Learning Theory*, pages 4–22, 2015. Cited on pages: 2, 16

Han Bao. Proper losses, moduli of convexity, and surrogate regret bounds. In *Conference on Learning Theory*, pages 525–547, 2023. Cited on page: 16

Han Bao and Asuka Takatsu. Proper losses regret at least 1/2-order. *Journal of Machine Learning Research*, 26(288):1–50, 2025. Cited on page: 16

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008. Cited on page: 2

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with Fenchel–Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020. Cited on page: 16

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, 2005. Cited on page: 16

Yuzhou Cao, Tianchi Cai, Lei Feng, Lihong Gu, Jinjie Gu, Bo An, Gang Niu, and Masashi Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. *Advances in Neural Information Processing Systems*, 35:521–534, 2022. Cited on page: 2

Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. *Advances in Neural Information Processing Systems*, 36:38485–38503, 2023. Cited on page: 16

Yuzhou Cao, Han Bao, Lei Feng, and Bo An. Establishing linear surrogate regret bounds for convex smooth losses via convolutional Fenchel–Young losses. *Advances in Neural Information Processing Systems*, 38:33200–33237, 2025. Cited on page: 16

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pages 1507–1517. PMLR, 2021. Cited on page: 2

Mohammad-Amin Charusaie and Samira Samadi. A unifying post-processing framework for multi-objective learn-to-defer problems. *Advances in Neural Information Processing Systems*, 37: 23705–23755, 2024. Cited on page: 16

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning*, pages 2972–3005. PMLR, 2022. Cited on page: 16

Xin Cheng, Yuzhou Cao, Haobo Wang, Hongxin Wei, Bo An, and Lei Feng. Regression with cost-based rejection. *Advances in Neural Information Processing Systems*, 36:45172–45196, 2023. Cited on page: 16

Chao-Kong Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. Cited on page: 3

Andreas Christmann and Ingo Steinwart. How SVMs can estimate quantiles and the median. *Advances in Neural Information Processing Systems*, 20:305–312, 2007. Cited on pages: 4, 6

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82, 2016a. Cited on pages: 2, 16

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. *Advances in Neural Information Processing Systems*, 29:1660–1668, 2016b. Cited on page: 16

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Theory and algorithms for learning with rejection in binary classification. *Annals of Mathematics and Artificial Intelligence*, 92(2):277–315, 2023. Cited on pages: 2, 16

Corinna Cortes, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Optimized deferral for imbalanced settings. *arXiv preprint arXiv:2604.27723*, 2026. Cited on page: 16

Balázs Csikós, György Kiss, Konrad J. Swanepoel, and Pieter Oloff de Wet. Large antipodal families. *Periodica Mathematica Hungarica*, 58(2):129–138, 2009. Cited on page: 23

Ludwig Danzer and Branko Grünbaum. Über zwei Probleme bezüglich konvexer Körper von P. Erdös und von V.L. Klee. *Mathematische Zeitschrift*, 79:95–99, 1962. Cited on page: 23

Giulia DeSalvo, Clara Mohri, Mehryar Mohri, and Yutao Zhong. Budgeted multiple-expert deferral. *arXiv preprint arXiv:2510.26706*, 2025. Cited on page: 16

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. *Advances in Neural Information Processing Systems*, 32:10781–10791, 2019. Cited on page: 9

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. Embedding dimension of polyhedral losses. In *Conference on Learning Theory*, pages 1558–1585, 2020. Cited on pages: 1, 2, 7, 8, 9, 10, 16, and 23

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. Unifying lower bounds on prediction dimension of convex surrogates. *Advances in Neural Information Processing Systems*, 34:22046–22057, 2021. Cited on pages: 2, 3, 6, 16, and 23

Jessica Finocchiaro, Rafael Frongillo, and Enrique Nueve. The structured abstain problem and the Lovász hinge. In *Conference on Learning Theory*, pages 3718–3740, 2022. Cited on pages: 1, 9, and 16

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for the design and analysis of consistent polyhedral surrogates. *Journal of Machine Learning Research*, 25(63): 1–60, 2024. Cited on pages: 1, 4, 5, 6, 7, 9, and 10

Tobias Fissler and Johanna F. Ziegel. Higher order elicitability and Osband's principle. *The Annals of Statistics*, pages 1680–1707, 2016. Cited on page: 16

Rafael Frongillo and Ian A. Kash. On elicitation complexity. *Advances in Neural Information Processing Systems*, 28:3258–3266, 2015a. Cited on page: 16

Rafael Frongillo and Ian A. Kash. Vector-valued property elicitation. In *Conference on Learning Theory*, pages 710–727. PMLR, 2015b. Cited on page: 16

Benno Fuchssteiner and Wolfgang Lusky. *Convex Cones*, volume 56. Elsevier, 2011. Cited on page: 23

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. Cited on page: 16

Wittawat Jitkrittum, Neha Gupta, Aditya K. Menon, Harikrishna Narasimhan, Ankit Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? *Advances in Neural Information Processing Systems*, 36:9891–9906, 2023. Cited on page: 16

Drona Khurana, Anish Thilagar, Dhamma Kimpara, and Rafael Frongillo. Consistency conditions for differentiable surrogate losses. *Advances in Neural Information Processing Systems*, 38: 135394–135430, 2025. Cited on pages: 4, 16

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *International Conference on Learning Representations*, pages 1–15, 2015. Cited on page: 25

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978. Cited on page: 6

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. Cited on page: 25

Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008. Cited on page: 16

Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Mitigating underfitting in learning to defer with consistent losses. In *International Conference on Artificial Intelligence and Statistics*, pages 4816–4824. PMLR, 2024. Cited on page: 16

Shuqi Liu, Yuzhou Cao, Lei Feng, Bo An, and Luke Ong. When more experts hurt: Underfitting in multi-expert learning to defer. *arXiv preprint arXiv:2602.17144*, 2026. Cited on page: 16

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018. Cited on page: 16

Anqi Mao. *Theory and Algorithms for Learning with Multi-Class Abstention and Multi-Expert Deferral*. New York University, 2025. Cited on page: 16

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. *Advances in Neural Information Processing Systems*, 36:3578–3606, 2023a. Cited on page: 16

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Ranking with abstention. *arXiv preprint arXiv:2307.02035*, 2023b. Cited on page: 16

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *International Workshop on Combinatorial Image Analysis*, pages 107–135. Springer, 2024a. Cited on page: 16

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable $\mathcal{H}$-consistent and Bayes-consistent loss functions for learning to defer. *Advances in Neural Information Processing Systems*, 37:73638–73671, 2024b. Cited on page: 16

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. In *International Conference on Machine Learning*, pages 34738–34759. PMLR, 2024c. Cited on page: 16

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, pages 822–867, 2024d. Cited on pages: 2, 16

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable $\mathcal{H}$-consistency and strong guarantees for learning to defer. In *International Conference on Machine Learning*, pages 43035–43066, 2025. Cited on page: 16

Christopher Mohri, Daniel Andor, Eunsol Choi, Michael Collins, Anqi Mao, and Yutao Zhong. Learning to reject with a fixed predictor: Application to decontextualization. In *International Conference on Learning Representations*, 2023. Cited on page: 16

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. One-stage top-k learning-to-defer: Score-based surrogates with theoretical guarantees. *arXiv preprint arXiv:2505.10160*, 2025. Cited on page: 16

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Beyond augmented-action surrogates for multi-expert learning-to-defer. *arXiv preprint arXiv:2604.09414*, 2026a. Cited on page: 16

Yannis Montreuil, Hoang Duy Dang, Maxime Meyer, Lai Xing Ng, Axel Carlier, and Wei Tsang Ooi. Online learning-to-defer with varying experts. In *International Conference on Artificial Intelligence and Statistics*, pages 2701–2709, 2026b. Cited on page: 16

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020. Cited on page: 16

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International Conference on Artificial Intelligence and Statistics*, pages 10520–10545. PMLR, 2023. Cited on page: 16

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K. Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. *Advances in Neural Information Processing Systems*, 35:29292–29304, 2022. Cited on page: 16

Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. *Advances in Neural Information Processing Systems*, 32: 2582–2592, 2019. Cited on pages: 2, 16

Enrique Nueve, Bo Waggoner, Dhamma Kimpara, and Jessica Finocchiaro. Trading off consistency and dimensionality of convex surrogates for multiclass classification. *Advances in Neural Information Processing Systems*, 37:103005–103030, 2024. Cited on page: 16

Nastaran Okati, Abir De, and Manuel Rodriguez. Differentiable learning under triage. *Advances in Neural Information Processing Systems*, 34:9140–9151, 2021. Cited on page: 16

B.T. Polyak. *Introduction to Optimization*. Translations series in mathematics and engineering. Optimization Software, Publications Division, 1987. ISBN 9780911575149. Cited on pages: 6, 18

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019. Cited on page: 16

Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(14):1–45, 2016. Cited on pages: 3, 4, and 16

Harish G. Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12:530–554, 2018. Cited on pages: 1, 2, 4, 9, 10, and 23

Stefan Reichelstein and Kent Osband. Information-eliciting compensation schemes. *Journal of Public Economics*, 46:107–115, 1985. Cited on page: 16

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010. Cited on page: 16

Ralph T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997. Cited on page: 25

Ralph T. Rockafellar and Stan Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1-2):33–53, 2013. Cited on page: 6

Ralph T. Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. Cited on page: 18

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971. Cited on page: 16

Song-Qing Shen, Bin-Bin Yang, and Wei Gao. AUC optimization with a reject option. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5684–5691, 2020. Cited on page: 16

Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and identification of properties. In *Conference on Learning Theory*, pages 482–526. PMLR, 2014. Cited on pages: 2, 6

Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(5):111–130, 2010. Cited on page: 2

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, 2016. doi: 10.5244/C.30.87. Cited on page: 25

Ahmed Zaoui, Christophe Denis, and Mohamed Hebiri. Regression with reject option and application to KNN. *Advances in Neural Information Processing Systems*, 33:20073–20082, 2020. Cited on page: 16

Johanna F. Ziegel. Coherence and elicitability. *Mathematical Finance*, 26(4):901–918, 2016. Cited on page: 16

A Additional Review on Related Works

More related works on learning with rejection. Beyond the design of calibrated surrogate losses for CwR, several methods train the label predictor and the rejector separately in a two-stage manner [Mohri et al., 2023, Mao et al., 2024d, Mao, 2025]. While classification with rejection has been studied in both binary and multiclass settings [Cortes et al., 2023, 2016b,a, Ni et al., 2019, Mao, 2025], the idea of augmenting the prediction space with a rejection option has also been explored beyond classification, including ranking [Mao et al., 2023b], regression [Zaoui et al., 2020, Cheng et al., 2023], AUC optimization [Shen et al., 2020], and structured prediction [Finocchiaro et al., 2022].

Recent work has also studied learning to defer to experts [Madras et al., 2018, Raghu et al., 2019, Okati et al., 2021, Mozannar and Sontag, 2020], which extends the deterministic rejection-cost setting to a stochastic one, where the expected cost corresponds to the expert's risk. A growing body of work has investigated calibrated surrogate losses for learning to defer [Charusaie et al., 2022, Cao et al., 2023, Mao et al., 2023a, 2024a,c,b, 2025], as well as efficient surrogate designs in this setting [Mozannar et al., 2023, Montreuil et al., 2025, 2026a,b, Liu et al., 2024, 2026, Narasimhan et al., 2022, Jitkrittum et al., 2023, Charusaie and Samadi, 2024, DeSalvo et al., 2025, Cortes et al., 2026].

Property elicitation and loss design. Property elicitation [Savage, 1971, Reichelstein and Osband, 1985, Lambert et al., 2008, Gneiting and Raftery, 2007, Frongillo and Kash, 2015b, Abernethy and Frongillo, 2012, Ziegel, 2016, Fissler and Ziegel, 2016] has played a central role in the design of surrogate losses. A canonical example is proper losses [Buja et al., 2005, Reid and Williamson, 2010, Bao, 2023, Bao and Takatsu, 2025], which elicit class probabilities and predict the label with the largest confidence. Other losses capable of eliciting class probabilities have also been widely used [Blondel et al., 2020, Cao et al., 2025]. More generally, Agarwal and Agarwal [2015] develop a principled framework for constructing calibrated surrogate losses from property-eliciting losses, showing that the elicitation of linear properties yields calibration under certain conditions. Further sufficient conditions connecting elicitation and calibration have recently been studied by Khurana et al. [2025].

Loss-dimension reduction is also closely tied to property elicitation. Ramaswamy and Agarwal [2016] initiated the study of convex calibration dimension, and Agarwal and Agarwal [2015] established lower bounds for calibrated losses based on the elicitation of linear properties. Lower bounds on the dimension required for property elicitation and their relation to convex calibration dimension have also been studied in Frongillo and Kash [2015a] and Finocchiaro et al. [2020, 2021]. Beyond assumption-free settings, lower-dimensional losses can be obtained under low-noise conditions, such as Massart noise, by carefully constructing lower-dimensional properties [Agarwal and Agarwal, 2015, Nueve et al., 2024].

B Proofs for Section 3

B.1 Proof of Lemma 7

Proof. According to the conditions in the lemma, suppose $\max_y \eta_y > 0.5$ and $y^* \coloneqq \operatorname{argmax}_y \eta_y$. We have

$$\Pr(Y < y^*) = \sum_{y=1}^{y^*-1} \eta_y \leq 1 - \eta_{y^*} < 0.5,$$

$$\Pr(Y \leq y^*) = \sum_{y=1}^{y^*} \eta_y \geq \eta_{y^*} > 0.5,$$

which concludes the proof. □

B.2 Proof of Lemma 8

Proof. Consider the case $\max_y \eta_y > 1-c$ and write $y^* = \operatorname{argmax}_y \eta_y$. Then we have $\Pr(|Y-y^*| < 0) = 0 < 1-c$ and $\Pr(|Y-y^*| \leq 0) = \eta_{y^*} > 1-c$, which implies $Q_{1-c}(|y^*-Y|) = \{0\}$.

Consider the case $\max_y \eta_y < 1 - c$. Fix an arbitrary $\alpha \in [0, 0.5)$. Then, we have for any $\theta \in \mathbb{R}$ that, if there does not exist $y' \in \mathcal{Y}$ such that $\theta \in [y' - \alpha, y' + \alpha]$ holds, then $\Pr(|Y - \theta| \leq \alpha) = 0 < 1 - c$ holds; conversely, for any $\theta \in \mathbb{R}$, if such y' exists, it is unique we have $\Pr(|Y - \theta| \leq \alpha) = \Pr(Y = y') \leq \max_y \eta_y < 1 - c$. This implies that any $\alpha < 0.5$ cannot be in $Q_{1-c}(|\theta - Y|)$, concluding $Q_{1-c}(|\theta - Y|) \geq 0.5$.

Consider the case $\max_y \eta_y = 1 - c$, for any θ, we have

- If $0.5 \leq \min_{y^* \in \mathrm{Argmax}_y \eta_y} |\theta - y^*|$, for any $\alpha \in [0, 0.5)$, we have that $|\theta - y^*| > \alpha$ for any $y^* \in \mathrm{Argmax}_y \eta_y$, and there exists at most one $y' \in \mathcal{Y} \setminus \mathrm{Argmax}_y \eta_y$ such that $|\theta - y'| \leq \alpha$. When such y' does not exist, $\Pr(|Y - \theta| \leq \alpha) = 0 < 1 - c$ naturally holds. When it exists, we have
$$\Pr(|Y - \theta| \leq \alpha) \leq \Pr(Y = y') = \eta_{y'} < \max_y \eta_y = 1 - c.$$

- If $0.5 > \min_{y^* \in \mathrm{Argmax}_y \eta_y} |\theta - y^*|$, for any $0 < \alpha < \min_{y^* \in \mathrm{Argmax}_y \eta_y} |\theta - y^*| < 0.5$, there is no $y \in \mathcal{Y}$ such that $|\theta - y| \leq \alpha$.

  This follows from a contradiction. Assume such y exists. If $y \in \mathrm{Argmax}_y \eta_y$, it contradicts $\alpha < \min_{y^* \in \mathrm{Argmax}_y \eta_y} |\theta - y^*|$. If $y \in \mathcal{Y} \setminus \mathrm{Argmax}_y \eta_y$, then we have $|\theta - y| < 0.5$ and $|\theta - y^*| = |(y^* - y) - (\theta - y)| \geq |y^* - y| - |\theta - y| \geq 1 - 0.5 = 0.5$, which is a contradiction.

These two observations indicate that any $\alpha < \min(0.5, \min_{y^* \in \mathrm{Argmax}_y \eta_y} |\theta - y^*|)$ is not in $Q_{1-c}(|\theta - Y|)$. □

B.3 Proof of Lemma 9

Proof. For any $\boldsymbol{\eta} \in \Delta^K$ and $[\theta, r]^\top \in \widetilde{\Gamma}(\boldsymbol{\eta})$:

- If $\max_y \eta_y > 1 - c$, we have $\gamma_{01c}(\boldsymbol{\eta}) = \{\mathrm{argmax}_y \eta_y\}$ and $\theta = \mathrm{argmax}_y \eta_y$ and $r = 0$ according to (3) and Lemmas 7, 8. Then $\varphi^\tau([\theta, r]^\top) = \mathrm{argmax}_y \eta_y \in \gamma_{01c}(\boldsymbol{\eta})$.
- If $\max_y \eta_y < 1 - c$, we have $\gamma_{01c}(\boldsymbol{\eta}) = \{®\}$, and $r \geq 0.5$ according to Lemma 8. Then $\varphi^\tau([\theta, r]^\top) = ® \in \gamma_{01c}(\boldsymbol{\eta})$.
- If $\max_y \eta_y = 1 - c$, then $\gamma_{01c}(\boldsymbol{\eta}) = \{®\} \cup \mathrm{Argmax}_y \eta_y$ and $r \geq \min(0.5, \min_{y^* \in \mathrm{Argmax}_y \eta_y} |\theta - y^*|)$ according to Lemma 8. When $\min_{y^* \in \mathrm{Argmax}_y \eta_y} |\theta - y^*| \geq 0.5$, we have $\varphi^\tau([\theta, r]^\top) = ® \in \gamma_{01c}(\boldsymbol{\eta})$. When $\min_{y^* \in \mathrm{Argmax}_y \eta_y} |\theta - y^*| < 0.5$, there exists $y^* \in \mathrm{Argmax}_y \eta_y$ such that $\widetilde{\mathrm{argmin}}_y(|\theta - y|) = y^*$. In this case $\varphi^\tau([\theta, r]^\top) \in \{®, y^*\} \subseteq \gamma_{01c}(\boldsymbol{\eta})$.

□

C Proofs for Section 4

C.1 Helper Lemmas

Before proving the main theoretical statements in Section 4, we first introduce several helper lemmas:

Lemma 26. *For $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, and consider the following two conditions:*

Condition 1: *$\inf_{\boldsymbol{y} \in \mathbb{R}^n} f(\boldsymbol{x}, \boldsymbol{y})$ is achieved by $\mathcal{S}_f(\boldsymbol{x}) \subseteq \mathbb{R}^n$ for any $\boldsymbol{x} \in \mathbb{R}^m$, i.e.,*
$$\inf_{\boldsymbol{y} \in \mathbb{R}^n} f(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y}'),\ \forall (\boldsymbol{x}, \boldsymbol{y}') \in \mathbb{R}^m \times \mathcal{S}_f(\boldsymbol{x}).$$

Condition 2: *Let $g(\boldsymbol{x}) \coloneqq \min_{\boldsymbol{y} \in \mathbb{R}^n} f(\boldsymbol{x}, \boldsymbol{y})$ and $g(\boldsymbol{x})$ is minimized by $\mathcal{G}_f$, i.e.,*
$$\inf_{\boldsymbol{x} \in \mathbb{R}^m} g(\boldsymbol{x}) = g(\boldsymbol{x}') = f(\boldsymbol{x}', \boldsymbol{y}'),\ \forall (\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{G}_f \times \mathcal{S}_f(\boldsymbol{x}').$$

Then $(\boldsymbol{x}, \boldsymbol{y})$ minimizes f if and only if $\boldsymbol{x} \in \mathcal{G}_f$ and $\boldsymbol{y} \in \mathcal{S}_f(\boldsymbol{x})$.

Proof. Suppose Condition 1 and 2 hold. For any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^m \times \mathbb{R}^n$, we have

$$f(\boldsymbol{x}, \boldsymbol{y}) \geq \inf_{\boldsymbol{y}'} f(\boldsymbol{x}, \boldsymbol{y}') \geq \inf_{\boldsymbol{x}'} \inf_{\boldsymbol{y}'} f(\boldsymbol{x}', \boldsymbol{y}') = \inf_{\boldsymbol{x}', \boldsymbol{y}'} f(\boldsymbol{x}', \boldsymbol{y}').$$

The equality of the first inequality holds if and only if $\boldsymbol{y} \in \mathcal{S}_f(\boldsymbol{x})$, and the equality of the second inequality holds if and only if $\boldsymbol{x} \in \mathcal{G}_f$, which concludes the proof. □

Here we focus on the following minimization problem, where X is a random variable.

$$F_\alpha(\boldsymbol{\theta}, r) = r + \frac{1}{1-\alpha}\mathbb{E}[\max(f(X, \boldsymbol{\theta}) - r, 0)], \tag{11}$$

and the following conclusion shows the connection between (11) and $\mathrm{CVaR}_\alpha(f(X, \boldsymbol{\theta}))$.

Lemma 27 (Equivalent Formula of CVaR [Rockafellar and Uryasev, 2002, Theorem 10])**.** *For $\alpha \in (0, 1)$, a function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ that is measurable in its first argument, continuous in its second argument, a random variable $X : \Omega \to \mathbb{R}^m$ that is independent of the second argument $\boldsymbol{\theta}$ of f, and $\mathbb{E}[|f(X, \boldsymbol{\theta})|] < +\infty$ for any $\boldsymbol{\theta} \in \mathbb{R}^n$, we have*

(a) $\mathrm{Argmin}_{r \in \mathbb{R}} F_\alpha(\boldsymbol{\theta}, r) = Q_\alpha(f(X, \boldsymbol{\theta}))$, $\forall \boldsymbol{\theta} \in \mathbb{R}^n$.

(b) $\mathrm{CVaR}_\alpha(f(X, \boldsymbol{\theta})) = \inf_{r \in \mathbb{R}} F_\alpha(\boldsymbol{\theta}, r)$, $\forall \boldsymbol{\theta} \in \mathbb{R}^n$.

The next lemma gives an explicit interval of the subgradient of pinball losses [Rockafellar and Uryasev, 2002].

Lemma 28. *For any random variable X with $\mathbb{E}[|X|] < \infty$ and $\alpha \in (0, 1)$, the subgradient of the expected pinball loss can be written as follows:*

$$\partial_\theta \mathbb{E}_X[\rho_\alpha(\theta, X)] = \left[\frac{\Pr(X < \theta) - \alpha}{1-\alpha}, \frac{\Pr(X \leq \theta) - \alpha}{1-\alpha}\right].$$

C.2 Proof of Theorem 10

Proof. First, since $R_{L^c}(\cdot, \boldsymbol{\eta})$ is polyhedral and non-negative for a fixed $\boldsymbol{\eta}$, it follows that $\mathrm{prop}[L^c]$ is non-empty for every $\boldsymbol{\eta} \in \Delta^K$. Lemma 27 implies that $\mathrm{Argmin}_r R_{L^c}([\theta, r]^\top, \boldsymbol{\eta}) = Q_{1-c}(|\theta - Y|)$ for all $\theta \in \mathbb{R}$ and $\boldsymbol{\eta} \in \Delta^K$, and $\widetilde{R}_{L^c}(\theta, \boldsymbol{\eta}) = \min_r R_{L^c}([\theta, r]^\top, \boldsymbol{\eta}) = \mathrm{CVaR}_{1-c}(|\theta - Y|) + \lambda \mathbb{E}_Y[|\theta - Y|]$. Since $\widetilde{R}_{L^c}(\theta, \boldsymbol{\eta})$ is also convex polyhedral w.r.t. θ, $\mathrm{Argmin}_\theta \widetilde{R}_{L^c}(\theta, \boldsymbol{\eta})$ is non-empty for each $\boldsymbol{\eta}$. According to Lemma 26, we have

$$\mathrm{Argmin}_{[\theta, r]^\top} R_{L^c}([\theta, r]^\top, \boldsymbol{\eta}) = \{[\theta, r]^\top \in \mathbb{R}^2 : \theta \in \mathrm{Argmin}_\theta \widetilde{R}_{L^c}(\theta, \boldsymbol{\eta}),\ r \in Q_{1-c}(|\theta - Y|)\}.$$

Then according to the definition of $\widetilde{\Gamma}$ (3), we need to prove the following claim (and another side $\max_{y \in \mathcal{Y}} \eta_y \leq 1 - c$ directly follows from the analyses above and the definition of $\widetilde{\Gamma}$):

$$\mathrm{Argmin}_\theta \widetilde{R}_{L^c}(\theta, \boldsymbol{\eta}) \subseteq \mathrm{Med}(Y) = \{y^* \coloneqq \mathrm{argmax}_y \eta_y\},\ \forall \boldsymbol{\eta} : \max_{y \in \mathcal{Y}} \eta_y > 1 - c.$$

Then we focus on subgradient analysis of $\widetilde{R}_{L^c}$. We have $\mathrm{CVaR}_{1-c}(|\theta - Y|)$ is 1-Lipschitz, and $\partial_\theta \mathrm{CVaR}_{1-c}(|\theta - Y|) + \lambda \partial_\theta \mathbb{E}[|\theta - Y|] \subseteq \partial_\theta \widetilde{R}_{L^c}(\theta, \boldsymbol{\eta})$. Then we have the following conclusion based on the observations above and Lemma 28:

$$\left[1 + \lambda \frac{\Pr(Y < y^*) - 0.5}{1 - 0.5}, -1 + \lambda \frac{\Pr(Y \leq y^*) - 0.5}{1 - 0.5}\right]$$
$$= [\lambda(2\Pr(Y < y^*) - 1) + 1, \lambda(2\Pr(Y \leq y^*) - 1) - 1] \subseteq \partial_{\theta = y^*} \widetilde{R}_{L^c}(\theta, \boldsymbol{\eta}).$$

Since $\Pr(Y < y^*) \leq 1 - (1 - c) = c$ and $\Pr(Y \leq y^*) \geq \eta_{y^*} \geq 1 - c$, we have

$$[1 - \lambda(1 - 2c), \lambda(1 - 2c) - 1] \subseteq \partial_{\theta = y^*} \widetilde{R}_{L^c}(\theta, \boldsymbol{\eta}).$$

Then we have that $0 \in \mathrm{int}([1 - \lambda(1 - 2c), \lambda(1 - 2c) - 1])$ if and only if $\lambda > \frac{1}{1-2c}$. In this case, $0 \in \mathrm{int}(\partial_{\theta = y^*} \widetilde{R}_{L^c}(\theta, \boldsymbol{\eta}))$ and thus y^* is the unique minimizer of $\widetilde{R}_{L^c}(\theta, \boldsymbol{\eta}))$ [Polyak, 1987, Lemma 5.3.3], which concludes the proof. □

Remark 29. *While Theorem 10 only requires the conclusion to hold for $\max_{y \in \mathcal{Y}} \eta_y > 1 - c$, the discussions above also hold when $\max_{y \in \mathcal{Y}} \eta_y = 1 - c$, which is crucial in the proof of Lemma 13.*

C.3 Proof of Lemma 13

Proof. When $\max_y \eta_y > 1 - c$, we have $\gamma_{01c}(\eta) = \{y^*\} \coloneqq \{\mathrm{argmax}_y \eta_y\}$. Then, for any $[\theta, r]^\top \in \mathbb{R}^2$ such that $\varphi^\tau([\theta, r]^\top) \notin \{y^*\}$, we have $|\theta - y^*| \geq 0.5$ or $r > \tau$ by definition of φ^τ in (4). According to Theorem 10, Lemma 7, Lemma 8, and the definition of $\widetilde{\Gamma}$ in (3), $\theta^* = y^*$ and $r^* = 0$ for any $[\theta^*, r^*]^\top \in \mathrm{prop}[L^c](\boldsymbol{\eta})$, thus $\|[\theta, r]^\top - [\theta^*, r^*]^\top\|_\infty = \max\{|\theta - \theta^*|, |r - r^*|\} \geq \tau$ for any $[\theta^*, r^*]^\top \in \mathrm{prop}[L^c](\boldsymbol{\eta})$.

When $\max_y \eta_y = 1 - c$, we have $\gamma_{01c}(\boldsymbol{\eta}) = \{y^*, ®\} \coloneqq \{\mathrm{argmax}_y \eta_y, ®\}$. Then, for any $[\theta, r]^\top \in \mathbb{R}^2$ such that $\varphi^\tau([\theta, r]^\top) \notin \{y^*, ®\}$, we have that $|\theta - y^*| \geq 0.5$ and $r \leq \tau$ by definition of φ^τ in (4). According to the proof of Theorem 10, $\theta^* = y^*$, then we have

$$\|[\theta, r]^\top - [\theta^*, r^*]^\top\|_\infty \geq |\theta - \theta^*| = |\theta - y^*| \geq 0.5.$$

When $\max_y \eta_y < 1 - c$, we have $\gamma_{01c}(\boldsymbol{\eta}) = \{®\}$. Then, for any $[\theta, r]^\top \in \mathbb{R}^2$ such that $\varphi^\tau([\theta, r]^\top) \notin \{®\}$, $r \leq \tau$ by definition of φ^τ in (4). According to Theorem 10, Lemma 8, and (3), $r^* \geq 0.5$ for any $[\theta^*, r^*]^\top \in \Gamma(\boldsymbol{\eta})$, thus $\|[\theta, r]^\top - [\theta^*, r^*]^\top\|_\infty \geq |r - r^*| \geq 0.5 - \tau$ for any $[\theta^*, r^*]^\top \in \mathrm{prop}[L^c](\boldsymbol{\eta})$.

Combining these cases, we have that for any $\boldsymbol{\eta} \in \Delta^K$, $\|[\theta, r]^\top - [\theta^*, r^*]^\top\|_\infty \geq \min(\tau, 0.5 - \tau)$ for any $[\theta^*, r^*]^\top \in \mathrm{prop}[L^c](\boldsymbol{\eta})$ and $[\theta, r]^\top \in \mathbb{R}^2$ such that $\varphi^\tau([\theta, r]^\top) \notin \gamma_{01c}(\boldsymbol{\eta})$, which indicates $\inf_{\boldsymbol{d} \in \mathrm{prop}[L^c](\boldsymbol{\eta})} \|[\theta, r]^\top - \boldsymbol{d}\|_\infty \geq \min(\tau, 0.5 - \tau)$. □

D Proof of Theorem 16

Proof. According to the definition of property, $\mathrm{prop}[L](\boldsymbol{\eta})$ is non-empty. Suppose there is a 2-dimensional convex loss L that $\mathrm{prop}[L](\boldsymbol{\eta}) \subseteq \widetilde{\Gamma}(\boldsymbol{\eta})$ for any $\boldsymbol{\eta} \in \Delta^K$. Let $\bar{\boldsymbol{\eta}} = [0.5, 0.5, 0, \cdots, 0]^\top \in \Delta^K$. For a sequence $\epsilon_n \downarrow 0$, define the following two sequences

$$\boldsymbol{\eta}^{-,n} = [0.5 + \epsilon_n, 0.5 - \epsilon_n, 0, \cdots, 0],$$
$$\boldsymbol{\eta}^{+,n} = [0.5 - \epsilon_n, 0.5 + \epsilon_n, 0, \cdots, 0].$$

Then, both $\boldsymbol{\eta}^{-,n}$ and $\boldsymbol{\eta}^{+,n}$ converge to $\bar{\boldsymbol{\eta}}$ as $n \to \infty$. For finite n, we have $\max_y \eta_y^{-,n} > 0.5$, and $\max_y \eta_y^{+,n} > 0.5$.

We show $[1, 0]^\top \in \mathrm{prop}[L](\bar{\eta})$ from now. To do so, consider $Y \sim \boldsymbol{\eta}^{-,n}$. We have $\mathrm{Med}(Y) = \{1\}$ since $\Pr(Y < 1) = 0$ and $\Pr(Y \leq 1) > 0.5$. Meanwhile, note that $c = 0.5$, $Q_{1-c}(|1 - Y|) = Q_{0.5}(|1 - Y|) = \mathrm{Med}(|1 - Y|) = \{0\}$ since $\Pr(|1 - Y| < 0) = 0 < 0.5$ and $\Pr(|1 - Y| \leq 0) > 0.5$. As a result, $\mathrm{prop}[L](\boldsymbol{\eta}^{-,n}) \subseteq \widetilde{\Gamma}(\boldsymbol{\eta}^{-,n}) = \{[1, 0]^\top\}$, which implies $\mathrm{prop}[L](\boldsymbol{\eta}^{-,n}) = \{[1, 0]^\top\}$ since the property is non-empty. For any $n > 0$ and $[\theta, r]^\top \in \mathbb{R}^2$, we have $R_L([1, 0]^\top, \boldsymbol{\eta}^{-,n}) \leq R_L([\theta, r]^\top, \boldsymbol{\eta}^{-,n})$ since $[1, 0]^\top$ minimizes the risk. Since $R_L([\theta, r]^\top, \boldsymbol{\eta}^{-,n})$ is linear w.r.t. $\boldsymbol{\eta}^{-,n}$, we have $R_L([1, 0]^\top, \bar{\boldsymbol{\eta}}) \leq R_L([\theta, r]^\top, \bar{\boldsymbol{\eta}})$ given $\boldsymbol{\eta}^{-,n}$ converges to $\bar{\boldsymbol{\eta}}$ as $n \to \infty$, which indicates $[1, 0]^\top \in \mathrm{prop}[L](\bar{\boldsymbol{\eta}})$.

Applying the same argument to the sequence $\boldsymbol{\eta}^{+,n}$, we obtain $[2, 0]^\top \in \mathrm{prop}[L](\bar{\boldsymbol{\eta}})$.

Note that $R_L([\theta, r]^\top, \bar{\boldsymbol{\eta}})$ is convex w.r.t. $[\theta, r]^\top$ since L is convex, which indicates that $\mathrm{prop}[L](\bar{\boldsymbol{\eta}})$ is also convex. This implies that $[1.5, 0]^\top = \frac{1}{2}[1, 0]^\top + \frac{1}{2}[2, 0]^\top$ is also in $\mathrm{prop}[L](\bar{\boldsymbol{\eta}})$. However, note that we have $\Pr(|1.5 - Y| < 0) = 0$ and $\Pr(|1.5 - Y| \leq 0) = 0$ when $Y \sim \bar{\boldsymbol{\eta}}$, which indicates $[1.5, 0]^\top \notin \widetilde{\Gamma}(\bar{\boldsymbol{\eta}})$ and thus not in $\mathrm{prop}[L](\bar{\boldsymbol{\eta}})$, which contradicts the assumption on the existence of L. □

E Proofs for Section 6

E.1 Proof of Lemma 17

Proof. When the weak dominant-label condition holds, for any $y^* \in \mathrm{Argmax}_y \eta_y$, we have $\eta_{y^*} \geq 0.5$. We first show that $y^* \notin Q_{0.5-\delta}(Y) \cup Q_{0.5+\delta}(Y)$ is impossible with proof by contradiction. When $y^* \notin Q_{0.5-\delta}(Y)$, we have

$$\Pr(Y < y^*) > 0.5 - \delta \quad \text{or} \quad \Pr(Y \leq y^*) < 0.5 - \delta.$$

Noting the weak dominant-label condition $\eta_{y^*} \geq 0.5$, we have $\Pr(Y < y^*) > 0.5 - \delta$ exclusively. When $y^* \notin Q_{0.5+\delta}(Y)$, we have

$$\Pr(Y < y^*) > 0.5 + \delta \quad \text{or} \quad \Pr(Y \leq y^*) < 0.5 + \delta.$$

Noting the weak dominant-label condition $\eta_{y^*} \geq 0.5$, we have $\Pr(Y \leq y^*) < 0.5 + \delta$ exclusively. Then $y^* \notin Q_{0.5-\delta}(Y) \cup Q_{0.5+\delta}(Y)$, we have $\Pr(Y < y^*) > 0.5 - \delta$ and $\Pr(Y \leq y^*) < 0.5 + \delta$. However, since $\Pr(Y \leq y^*) = \Pr(Y < y^*) + \eta_{y^*} > 0.5 - \delta + 0.5 = 1 - \delta > 0.5 + \delta$ for $\delta \in (0, 0.25)$, this leads to a contradiction, and we have $y^* \in Q_{0.5-\delta}(Y) \cup Q_{0.5+\delta}(Y)$.

Then we further prove that $Q_{0.5-\delta}(Y) = \{y^*\}$ or $Q_{0.5+\delta}(Y) = \{y^*\}$.

- If $y^* \in Q_{0.5+\delta}(Y)$ but $Q_{0.5+\delta}(Y) \neq \{y^*\}$, we have $\Pr(Y \leq y^*) = 0.5 + \delta$ (otherwise $Q_{0.5+\delta}(Y) = \{y^*\}$). In this case, $\Pr(Y < y^*) \leq \delta < 0.5 - \delta$ and $\Pr(Y \leq y^*) \geq 0.5 + \delta > 0.5 - \delta$, and thus $Q_{0.5-\delta}(Y) = \{y^*\}$ since $\delta \in (0, 0.25)$.
- If $y^* \in Q_{0.5-\delta}(Y)$ but $Q_{0.5-\delta}(Y) \neq \{y^*\}$, we have $\Pr(Y < y^*) = 0.5 - \delta$ (otherwise $Q_{0.5-\delta}(Y) = \{y^*\}$). In this case, $\Pr(Y < y^*) = 0.5 - \delta < 0.5 + \delta$ and $\Pr(Y \leq y^*) = 1 - \delta > 0.5 + \delta$, and thus $Q_{0.5+\delta}(Y) = \{y^*\}$ since $\delta \in (0, 0.25)$.

Combining the discussions above, we conclude the proof. □

E.2 Proof of Theorem 18

We break the proof of this theorem into two parts.

Lemma 30 (Minimizer Analysis-θ)**.** *Set $\delta \in (0, 0.25)$, $\lambda^+ > \frac{0.5-\delta}{\min(\delta, 0.25-\delta)}$, and $\lambda^- > \frac{0.5+\delta}{\min(0.25-\delta, \delta)}$:*

- *If $\max_y \eta_y > 0.5$ holds, then we have either $\operatorname{Argmin}_\theta \min_r R_{L^+}([\theta, r]^\top, \boldsymbol{\eta}) = \{y^*\}$ or $\operatorname{Argmin}_\theta \min_r R_{L^-}([\theta, r]^\top, \boldsymbol{\eta}) = \{y^*\}$, where $y^* = \operatorname{argmax}_y \eta_y$.*
- *If $\max_y \eta_y = 0.5$ holds, then we have either $\operatorname{Argmin}_\theta \min_r R_{L^+}([\theta, r]^\top, \boldsymbol{\eta}) = \{y'\}$ or $\operatorname{Argmin}_\theta \min_r R_{L^-}([\theta, r]^\top, \boldsymbol{\eta}) = \{y'\}$, where $y' \in \operatorname{Argmax}_y \eta_y$.*

Proof. According to the definition of $\widetilde{R}_{L^+}$ and $\widetilde{R}_{L^-}$, $\widetilde{R}_{L^+}(\theta, \boldsymbol{\eta}) = \min_r R_{L^+}([\theta, r]^\top, \boldsymbol{\eta}) = \mathrm{CVaR}_{0.5}(|Y-\theta|) + \lambda^+ \mathbb{E}_Y[\rho_{0.5+\delta}(\theta, Y)]$ and $\widetilde{R}_{L^-}(\theta, \boldsymbol{\eta}) = \min_r R_{L^-}([\theta, r]^\top, \boldsymbol{\eta}) = \mathrm{CVaR}_{0.5}(|Y-\theta|) + \lambda^- \mathbb{E}_Y[\rho_{0.5-\delta}(\theta, Y)]$. Note that $\mathrm{CVaR}_{0.5}(|Y-\theta|)$ is 1-Lipschitz. According to Lemma 28, for any (fixed) $y' \in \operatorname{Argmax}_y \eta_y$, we have

$$\partial_{\theta=y'} \mathrm{CVaR}_{0.5}(|\theta - Y|) + \lambda^{\pm} \partial_{\theta=y'} \mathbb{E}_Y[\rho_{0.5\pm\delta}(\theta, Y)] \subseteq \partial_{\theta=y'} \widetilde{R}_{L^\pm}(\theta, \boldsymbol{\eta}),$$

and according to Lemma 28 we have

$$\underbrace{\left[1 + \lambda^+ \frac{\Pr(Y < y') - (0.5+\delta)}{1 - (0.5+\delta)}, -1 + \lambda^+ \frac{\Pr(Y \leq y') - (0.5+\delta)}{1 - (0.5+\delta)}\right]}_{\mathcal{I}_1} \subseteq \partial_{\theta=y'} \widetilde{R}_{L^+}(\theta, \boldsymbol{\eta}),$$

$$\underbrace{\left[1 + \lambda^- \frac{\Pr(Y < y') - (0.5-\delta)}{1 - (0.5-\delta)}, -1 + \lambda^- \frac{\Pr(Y \leq y') - (0.5-\delta)}{1 - (0.5-\delta)}\right]}_{\mathcal{I}_2} \subseteq \partial_{\theta=y'} \widetilde{R}_{L^-}(\theta, \boldsymbol{\eta}).$$

Given $\eta_{y'} \geq 0.5$, then $\Pr(Y < y') \leq 0.5$ and $\Pr(Y \leq y') \geq 0.5$, we focus on making $0 \in \operatorname{int}(\mathcal{I}_1)$ or $\operatorname{int}(\mathcal{I}_2)$ and hence $\theta = y'$ is a sharp and thus strict minima. To make $0 \in \operatorname{int}(\mathcal{I}_1) \cup \operatorname{int}(\mathcal{I}_2)$, we need that for any $\boldsymbol{\eta}$ such that $\max_y \eta_y \geq 0.5$,

$$1 + \lambda^+ \frac{\Pr(Y < y') - (0.5+\delta)}{1 - (0.5+\delta)} < 0, \; -1 + \lambda^+ \frac{\Pr(Y \leq y') - (0.5+\delta)}{1 - (0.5+\delta)} > 0,$$

or

$$1 + \lambda^- \frac{\Pr(Y < y') - (0.5-\delta)}{1 - (0.5-\delta)} < 0, \; -1 + \lambda^- \frac{\Pr(Y \leq y') - (0.5-\delta)}{1 - (0.5-\delta)} > 0.$$

According to the range of $\Pr(Y < y')$ and $\Pr(Y \le y')$, it is sufficient to fulfill

$$1 + \lambda^+ \frac{0.5 - (0.5+\delta)}{1-(0.5+\delta)} < 0,\ -1 + \lambda^+ \frac{\Pr(Y \le y') - (0.5+\delta)}{1-(0.5+\delta)} > 0, \tag{12}$$

or

$$1 + \lambda^- \frac{\Pr(Y < y') - (0.5-\delta)}{1-(0.5-\delta)} < 0,\ -1 + \lambda^- \frac{0.5 - (0.5-\delta)}{1-(0.5-\delta)} > 0. \tag{13}$$

The first inequality of (12) and the right inequality of (13) indicate that $\lambda^+ > \frac{0.5-\delta}{\delta}$ and $\lambda^- > \frac{0.5+\delta}{\delta}$. Furthermore, we have

- If $\Pr(Y < y') \le 0.25$, we have $\frac{\Pr(Y<y')-(0.5-\delta)}{1-(0.5-\delta)} \le \frac{\delta-0.25}{0.5+\delta} < 0$. Then it follows from the first inequality of (13) that

$$1 + \lambda^- \underbrace{\frac{\Pr(Y < y') - (0.5-\delta)}{1-(0.5-\delta)}}_{<0} < 0 \implies \lambda^- > \frac{0.5+\delta}{0.5-\delta-\Pr(Y<y')}.$$

  Since we require that this inequality holds for all $\Pr(Y < y') \le 0.25$, we further obtain the range of $\lambda^- > \frac{0.5+\delta}{0.25-\delta}$.

- If $\Pr(Y < y') \ge 0.25$, we have $\frac{\Pr(Y\le y')-(0.5+\delta)}{1-(0.5+\delta)} = \frac{\Pr(Y<y')+\eta_{y'}-(0.5+\delta)}{0.5-\delta} > \frac{0.75-(0.5+\delta)}{0.5-\delta} = \frac{0.25-\delta}{0.5-\delta} > 0$. According to the second inequality of (12), we have

$$-1 + \lambda^+ \underbrace{\frac{\Pr(Y \le y') - (0.5+\delta)}{1-(0.5+\delta)}}_{>0} > 0 \implies \lambda^+ > \frac{0.5-\delta}{\Pr(Y\le y') - (0.5+\delta)}.$$

  Since we require that this inequality holds for all $\Pr(Y < y') \ge 0.25$, given $\Pr(Y = y') \ge 0.5$, and thus $\Pr(Y \le y') \ge 0.75$. Then we further obtain the range of $\lambda^+ > \frac{0.5-\delta}{0.25-\delta}$.

Combining all these cases, we conclude the proof of the case of $\max_y \eta_y = 0.5$. Since $y' = y^*$ when $\max_y \eta_y > 0.5$, we conclude the proof for the case of $\max_y \eta_y > 0.5$. □

Lemma 31 (Minimizer Analysis-r)**.** *Set $\delta \in (0, 0.25)$, $\lambda^+ > \frac{0.5-\delta}{\min(\delta, 0.25-\delta)}$, and $\lambda^- > \frac{0.5+\delta}{\min(0.25-\delta,\delta)}$. Then $\mathrm{prop}[L^+](\boldsymbol{\eta})$ and $\mathrm{prop}[L^-](\boldsymbol{\eta})$ are non-empty. For any $[\theta^*, r^*]^\top \in \mathrm{prop}[L^+](\boldsymbol{\eta})$ or $\in \mathrm{prop}[L^-](\boldsymbol{\eta})$:*

- *If $\max_y \eta_y < 0.5$, we have $r^* \ge 0.5$.*
- *If $\max_y \eta_y > 0.5$, we have $r^* = 0$ if $\theta^* = \mathrm{argmax}_y \eta_y$.*
- *If $\max_y \eta_y = 0.5$, we have $r^* \ge \min(0.5, \min_{y' \in \mathrm{Argmax}_y \eta_y} |\theta^* - y'|)$.*

Proof. The non-emptiness follows directly from the convex polyhedral and non-negative nature of L^+ and L^-. According to Lemma 26, we have $r^* \in Q_{0.5}(|Y - \theta^*|) = \mathrm{Median}(|Y - \theta^*|)$. Then set $c = 0.5$ and we conclude the proof with Lemma 8. □

Then we begin the proof of Theorem 18.

Proof. The non-empty nature of $\mathrm{prop}[\widehat{L}]$ follows from the fact that $\widehat{L}$ is convex polyhedral and non-negative. Since for any $[\theta_1^*, r_1^*, \theta_2^*, r_2^*]^\top \in \mathrm{prop}[\widehat{L}](\boldsymbol{\eta})$, we have $[\theta_1^*, r_1^*]^\top \in \mathrm{prop}[L^+](\boldsymbol{\eta})$ and $[\theta_2^*, r_2^*]^\top \in \mathrm{prop}[L^-](\boldsymbol{\eta})$, we complete the proof with the two lemmas above. □

E.3 Proof of Theorem 19

Proof. For $[\theta_1^*, r_1^*, \theta_2^*, r_2^*] \in \mathrm{prop}[\widehat{L}](\boldsymbol{\eta})$, we have

- If $\max_y \eta_y < 0.5$, $\gamma_{01c}(\boldsymbol{\eta}) = \{®\}$ for $c = 0.5$, $r_1^*, r_2^* \geq 0.5$. For any $[\theta_1, r_1, \theta_2, r_2]^\top \in \mathbb{R}^4$ such that $\hat{\varphi}^\tau([\theta_1, r_1, \theta_2, r_2]^\top) \neq ®$, we have that $r_1 \leq \tau$ or $r_2 \leq \tau$. Thus $\|[\theta_1, r_1, \theta_2, r_2]^\top - [\theta_1^*, r_1^*, \theta_2^*, r_2^*]^\top\|_\infty \geq \max(|r_1 - r_1^*|, |r_2 - r_2^*|) \geq 0.5 - \tau$.
- If $\max_y \eta_y = 0.5$, $\gamma_{01c}(\boldsymbol{\eta}) = \{®\} \cup \mathrm{Argmax}_y \eta_y$, $r_i^* \geq \min(0.5, \min_{y' \in \mathrm{Argmax}_y \eta_y} |\theta_i^* - y'|)$ for $i \in \{1, 2\}$. For any $[\theta_1, r_1, \theta_2, r_2]^\top \in \mathbb{R}^4$ such that $\hat{\varphi}^\tau([\theta_1, r_1, \theta_2, r_2]^\top) \notin \gamma_{01c}(\boldsymbol{\eta})$, $r_1 \leq \tau$ and $\min_{y' \in \mathrm{Argmax}_y \eta_y} |\theta_1 - y'| \geq 0.5$ or $r_2 \leq \tau$ and $\min_{y' \in \mathrm{Argmax}_y \eta_y} |\theta_2 - y'| \geq 0.5$. Suppose $r_i \leq \tau$ and $\min_{y' \in \mathrm{Argmax}_y \eta_y} |\theta_i - y'| \geq 0.5$. Thus for any $y' \in \mathrm{Argmin}_{y' \in \mathrm{Argmax}_y \eta_y} |\theta_i^* - y'|$ we have

$$\begin{aligned}
&\|[\theta_1, r_1, \theta_2, r_2]^\top - [\theta_1^*, r_1^*, \theta_2^*, r_2^*]^\top\|_\infty \geq \max(|\theta_i - \theta_i^*|, |r_i - r_i^*|) \\
&\geq \max(|\theta_i - y'| - |\theta_i^* - y'|, \min(|\theta_i^* - y'|, 0.5) - \tau) \\
&\geq \max(0.5 - |\theta_i^* - y'|, \min(|\theta_i^* - y'|, 0.5) - \tau) \\
&\geq \frac{0.5 - \tau}{2}.
\end{aligned}$$

- If $\max_y \eta_y > 0.5$, $\gamma_{01c}(\boldsymbol{\eta}) = \{y^*\}$, $r_1^* = 0$ and $\theta_1^* = y^*$ or $r_2^* = 0$ and $\theta_2^* = y^*$. For any $[\theta_1, r_1, \theta_2, r_2]^\top \in \mathbb{R}^4$ such that $\hat{\varphi}^\tau([\theta_1, r_1, \theta_2, r_2]^\top) \neq y^*$, $r_1, r_2 > \tau$ or $r_1 \leq \tau$ and $|\theta_1 - y^*| \geq 0.5$ or $r_2 \leq \tau$ and $|\theta_2 - y^*| \geq 0.5$. Thus $\|[\theta_1, r_1, \theta_2, r_2]^\top - [\theta_1^*, r_1^*, \theta_2^*, r_2^*]^\top\|_\infty = \max(|\theta_1 - \theta_1^*|, |\theta_2 - \theta_2^*|, |r_1 - r_1^*|, |r_2 - r_2^*|)$.
  - If $r_1, r_2 > \tau$, the distance is lower bounded by τ.
  - Suppose $r_i \leq \tau$ and $|\theta_i - y^*| \geq 0.5$. Then we have

$$\begin{aligned}
&\max(|\theta_1 - \theta_1^*|, |\theta_2 - \theta_2^*|, |r_1 - r_1^*|, |r_2 - r_2^*|) \geq \max(|\theta_i - \theta_i^*|, |r_i^* - r_i|) \\
&\geq \max(|\theta_i - y^*| - |\theta_i^* - y^*|, \min(0.5, |\theta_i^* - y^*|) - \tau) \geq \frac{0.5 - \tau}{2}.
\end{aligned}$$

Combining these observations, we conclude the proof. □

F Proofs for Section 7

F.1 Proof of Theorem 23

Proof. Assume that there exists a convex surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ and an embedding $\boldsymbol{\psi}$ such that L embeds ℓ_{01c}, then we have $L(\boldsymbol{\psi}(®), y) = 0.5$ for all $y \in \mathcal{Y}$ and $L(\boldsymbol{\psi}(y'), y) = [\![y' \neq y]\!]$.

For any two different $i, j \in \mathcal{Y}$, let $\boldsymbol{\eta}_{ij} \in \Delta^K$ such that $\eta_i = \eta_j = 0.5$ and $\eta_k = 0$ for any $k \notin \{i, j\}$. According to Chow's rule (2), it follows that $®, i, j \in \mathrm{prop}[\ell_{01c}](\boldsymbol{\eta}_{ij}) = \gamma_{01c}(\boldsymbol{\eta}_{ij})$. Then from the definition of embedding, we have $\boldsymbol{\psi}(®), \boldsymbol{\psi}(i), \boldsymbol{\psi}(j) \in \mathrm{prop}[L](\boldsymbol{\eta}_{ij})$.

In this case, we have $R_L(\boldsymbol{\theta}, \boldsymbol{\eta}_{ij}) = \frac{1}{2} L(\boldsymbol{\theta}, i) + \frac{1}{2} L(\boldsymbol{\theta}, j)$. Furthermore, it follows from the loss value of $L(\boldsymbol{\psi}(\cdot), y)$ we analyzed above that

$$\underbrace{R_L(\boldsymbol{\psi}(i), \boldsymbol{\eta}_{ij})}_{\frac{1}{2}*0+\frac{1}{2}*1} = \underbrace{R_L(\boldsymbol{\psi}(j), \boldsymbol{\eta}_{ij})}_{\frac{1}{2}*1+\frac{1}{2}*0} = \underbrace{R_L(\boldsymbol{\psi}(®), \boldsymbol{\eta}_{ij})}_{\frac{1}{2}*0.5+\frac{1}{2}*0.5} = 0.5.$$

Since the embeddings of $i, j, ®$ are in $\mathrm{prop}[L](\boldsymbol{\eta}_{ij})$, i.e., $\boldsymbol{\psi}(i), \boldsymbol{\psi}(j), \boldsymbol{\psi}(®)$ minimize $R_L(\boldsymbol{\theta}, \boldsymbol{\eta}_{ij})$, we have $R_L(\boldsymbol{\theta}, \boldsymbol{\eta}_{ij}) \geq 0.5$ for any $\boldsymbol{\theta} \in \mathbb{R}^d$, which indicates

$$\frac{1}{2} L(\boldsymbol{\theta}, i) + \frac{1}{2} L(\boldsymbol{\theta}, j) \geq 0.5, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

It follows from the arbitrariness of i, j that

$$L(\boldsymbol{\theta}, i) + L(\boldsymbol{\theta}, j) \geq 1, \forall \boldsymbol{\theta} \in \mathbb{R}^d, \ i \neq j, i, j \in \mathcal{Y}.$$

Then

$$\underbrace{0.5 - L(\boldsymbol{\theta}, i)}_{\text{(A)}} \leq \underbrace{L(\boldsymbol{\theta}, j) - 0.5}_{\text{(B)}}, \forall \boldsymbol{\theta} \in \mathbb{R}^d,\ i \neq j, i, j \in \mathcal{Y}.$$

Since (A) is concave and (B) is convex, we have that there exists an affine function $f_{ij}(\boldsymbol{\theta})$ [Fuchssteiner and Lusky, 2011, p36] such that

$$0.5 - L(\boldsymbol{\theta}, i) \leq f_{ij}(\boldsymbol{\theta}) \leq L(\boldsymbol{\theta}, j) - 0.5, \forall \boldsymbol{\theta} \in \mathbb{R}^d,\ i \neq j, i, j \in \mathcal{Y}.$$

Given $L(\boldsymbol{\psi}(\circledR), i) = L(\boldsymbol{\psi}(\circledR), j) = 0.5$, we have

$$\underbrace{0.5 - L(\boldsymbol{\psi}(\circledR), i)}_{=0} \leq f_{ij}(\boldsymbol{\psi}(\circledR)) \leq \underbrace{L(\boldsymbol{\psi}(\circledR), j) - 0.5}_{=0},\ i \neq j, i, j \in \mathcal{Y}.$$

This means that $f_{ij}(\boldsymbol{\psi}(\circledR)) = 0$, which indicates that the affine function f_{ij} can be written as $f_{ij}(\boldsymbol{\theta}) \coloneqq \boldsymbol{\omega}_{ij}^\top(\boldsymbol{\theta} - \boldsymbol{\psi}(\circledR))$. Note that for any $i, j \in \mathcal{Y}$, $i \neq j$, and $k \in \mathcal{Y} \setminus \{i, j\}$, $L(\psi(k), i) = L(\psi(k), j) = 1$, Then

$$\underbrace{0.5 - L(\boldsymbol{\psi}(k), i)}_{=-0.5} \leq \boldsymbol{\omega}_{ij}^\top(\boldsymbol{\psi}(k) - \boldsymbol{\psi}(\circledR)) \leq \underbrace{L(\boldsymbol{\psi}(k), j) - 0.5}_{=0.5},$$

and since $L(\psi(j), i) = L(\psi(i), j) = 1$ and $L(\psi(i), i) = L(\psi(j), j) = 0$, we have

$$\underbrace{0.5 - L(\boldsymbol{\psi}(i), i)}_{=0.5} \leq \boldsymbol{\omega}_{ij}^\top(\boldsymbol{\psi}(i) - \boldsymbol{\psi}(\circledR)) \leq \underbrace{L(\boldsymbol{\psi}(i), j) - 0.5}_{0.5},$$

and

$$\underbrace{0.5 - L(\boldsymbol{\psi}(j), i)}_{=-0.5} \leq \boldsymbol{\omega}_{ij}^\top(\boldsymbol{\psi}(j) - \boldsymbol{\psi}(\circledR)) \leq \underbrace{L(\boldsymbol{\psi}(j), j) - 0.5}_{-0.5}.$$

Then it follows that for any $i, j \in \mathcal{Y}$, $i \neq j$, and $k \in \mathcal{Y} \setminus \{i, j\}$, there exists $\boldsymbol{\omega}_{ij}$ such that

$$\boldsymbol{\omega}_{ij}^\top(\boldsymbol{\psi}(i) - \boldsymbol{\psi}(\circledR)) = 0.5,\ \boldsymbol{\omega}_{ij}^\top(\boldsymbol{\psi}(j) - \boldsymbol{\psi}(\circledR)) = -0.5,\ \boldsymbol{\omega}_{ij}^\top(\boldsymbol{\psi}(k) - \boldsymbol{\psi}(\circledR)) \in [-0.5, 0.5].$$

According to Danzer and Grünbaum [1962, II a)] and Csikós et al. [2009], this indicates that the set $\{\boldsymbol{\psi}(y) - \boldsymbol{\psi}(\circledR)\}_{y=1}^K$ is an antipodal set, and thus $K \leq 2^d$, i.e., $d \geq \log_2 K$, which concludes the proof. □

F.2 Proof of Theorem 24

Proof. It follows from Theorem 23 that the embedding dimension of ℓ_{01c} when $c = 0.5$ is larger than 4 in this case, while the convex calibration dimension is no larger than 4 according to the calibration of $\widehat{L}$, which concludes the proof. □

G Equivalence of Constrained/Unconstrained Loss Dimensions in $\mathbb{R}$

Throughout this paper, we define the convex calibration dimension based on Definition 15, which is the minimally possible dimension d of the surrogate loss domain $\mathcal{C} \subseteq \mathbb{R}^d$ among any convex calibrated surrogate loss $L : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}$. Importantly, we explore all possible *constrained* surrogate losses. While this definition is the most general one and proposed in Ramaswamy et al. [2018], the follow-up work often considers the *unconstrained* surrogate losses $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ [Finocchiaro et al., 2020, 2021]. For our paper, we borrow the nonexistence result of 1-dimensional CwR surrogate losses [Finocchiaro et al., 2020, Theorem 12, Proposition 20] to establish Corollary 14, which supposes unconstrained $\mathcal{C} = \mathbb{R}^d$; thus we need to fill the gap between the constrained and unconstrained losses, strictly speaking. Below, we show there is no gap as long as we consider the 1-dimensional case, which legitimately justifies Corollary 14.

Theorem 32. *For a target loss $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, if there exist convex $\mathcal{C} \subseteq \mathbb{R}$, convex calibrated surrogate loss $L : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}$, and its link function $\varphi : \mathcal{C} \to \widehat{\mathcal{Y}}$, then there exist a loss $L' : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ and link $\varphi' : \mathbb{R} \to \widehat{\mathcal{Y}}$ such that (L', φ') is also calibrated with ℓ.*

G.1 Technical Lemma

To show Theorem 32, we need one technical lemma. We show it after preparing some notations. Let $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\infty\}$ be the extended real line. Recall $R_\ell(t, \boldsymbol{\eta}) \coloneqq \mathbb{E}_{Y\sim\boldsymbol{\eta}}[\ell(t, Y)]$ and $R_L(\theta, \boldsymbol{\eta}) \coloneqq \mathbb{E}_{Y\sim\boldsymbol{\eta}}[L(\theta, Y)]$, where we consider $t \in \widehat{\mathcal{Y}}$ and $\theta \in \mathcal{C} \subseteq \mathbb{R}$. The property $\gamma_\ell : \Delta^K \rightrightarrows \widehat{\mathcal{Y}}$ induced from ℓ is $\gamma_\ell(\boldsymbol{\eta}) \coloneqq \operatorname{Argmin}_{t\in\widehat{\mathcal{Y}}} R_\ell(t, \boldsymbol{\eta})$. For any non-empty $\mathcal{S} \subseteq \widehat{\mathcal{Y}}$, $\gamma_\mathcal{S} \coloneqq \{\boldsymbol{\eta} \in \Delta^K : \gamma_\ell(\boldsymbol{\eta}) = \mathcal{S}\}$ is a convex set because ℓ elicits γ and the minimizer of $R_\ell(\cdot, \boldsymbol{\eta})$ is closed under the convex combination of $\boldsymbol{\eta}$. The function $B(\boldsymbol{\eta}) \coloneqq \inf_{\theta\in\mathcal{C}, \varphi(\theta)\notin\gamma_\ell(\boldsymbol{\eta})} R_L(\theta, \boldsymbol{\eta})$ is the minimal risk of 'wrong' predictions. The set $\Theta^{\boldsymbol{\eta}} \coloneqq \{\theta \in \mathcal{C} : R_L(\theta, \boldsymbol{\eta}) < B(\boldsymbol{\eta})\}$ and $\Theta_\mathcal{S} \coloneqq \cup_{\boldsymbol{\eta}\in\gamma_\mathcal{S}} \Theta^{\boldsymbol{\eta}}$. Define $N_\epsilon(\theta) \coloneqq \{\theta' \in \mathbb{R} : |\theta - \theta'| < \epsilon\}$.

Lemma 33. *For a target loss $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, if there exists a convex non-empty and non-singleton $\mathcal{C} \subseteq \mathbb{R}$ with effective domain $\mathcal{C}$, a convex calibrated surrogate loss $L : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}$, and its link function $\varphi : \mathcal{C} \to \widehat{\mathcal{Y}}$, then there exists a closed interval $\mathcal{P} \coloneqq [a, b] \subseteq \operatorname{int}(\mathcal{C})$ such that $(L_{|\mathcal{P}}, \varphi)$ is also calibrated with ℓ, where $L_{|\mathcal{P}} : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$ is the restriction of L on $\mathcal{P}$.*

Proof. For every $\mathcal{S} \subseteq \widehat{\mathcal{Y}}$ such that $\gamma_\mathcal{S} \neq \emptyset$, select $\boldsymbol{\eta}_\mathcal{S} \in \gamma_\mathcal{S}$ and $\theta_\mathcal{S} \in \Theta^{\boldsymbol{\eta}_\mathcal{S}} \cap \operatorname{int}(\mathcal{C})$.[4] Given $P_\mathcal{S} = \{\mathcal{S} \subseteq \widehat{\mathcal{Y}} : \gamma_\mathcal{S} \neq \emptyset\}$ is finite, there exists $[a, b] \subseteq \operatorname{int}(\mathcal{C})$ such that $a < \theta_\mathcal{S} < b$ for any $\mathcal{S} \in P_\mathcal{S}$.

We complete the proof by contradiction. For any such interval $[a, b]$, assume $(L_{|[a,b]}, \varphi)$ is not calibrated, that is, there exists $\boldsymbol{\eta} \in \Delta^K$ such that[5]

$$\inf_{\theta\in[a,b], \varphi(\theta)\notin\gamma_\ell(\boldsymbol{\eta})} R_L(\theta, \boldsymbol{\eta}) \geq \min_{\theta\in[a,b]} R_L(\theta, \boldsymbol{\eta}) \geq \inf_{\theta\in[a,b], \varphi(\theta)\notin\gamma_\ell(\boldsymbol{\eta})} R_L(\theta, \boldsymbol{\eta}) \geq B(\boldsymbol{\eta}) > \inf_{\theta\in\mathcal{C}} R_L(\theta, \boldsymbol{\eta}), \tag{14}$$

where the second inequality follows from the fact that $(L_{|[a,b]}, \varphi)$ is not calibrated. The last inequality holds due to the assumption that L is calibrated. From the first two inequalities we have

$$\inf_{\theta\in[a,b], \varphi(\theta)\notin\gamma_\ell(\boldsymbol{\eta})} R_L(\theta, \boldsymbol{\eta}) = \min_{\theta\in[a,b]} R_L(\theta, \boldsymbol{\eta}). \tag{15}$$

According to the convexity and calibration of L, $R_L(\theta, \boldsymbol{\eta})$ is uniquely minimized at either $\theta = a$ or $\theta = b$; otherwise this minimizer would also minimize the risk over $\mathcal{C}$, which contradicts (14). Without loss of generality, assume this minimizer is b, and the case a is symmetric.

It follows from the convexity of $R_L(\cdot, \boldsymbol{\eta})$ and $\operatorname{argmin}_{\theta\in[a,b]} R_L(\theta, \boldsymbol{\eta}) = b$ that $R_L(\theta, \boldsymbol{\eta})$ is non-increasing on $\mathcal{C} \cap (-\infty, b]$. Denote by $\mathcal{S} \coloneqq \gamma_\ell(\boldsymbol{\eta})$. We have $\theta_\mathcal{S} < b$ by construction of the interval $[a, b]$, and then $R_L(\theta_\mathcal{S}, \boldsymbol{\eta}) > R_L(b, \boldsymbol{\eta})$. According to (15), there exists $u \in [a, b]$ such that

$$\varphi(u) \notin \mathcal{S}, \;\; R_L(u, \boldsymbol{\eta}) < R_L(\theta_\mathcal{S}, \boldsymbol{\eta}). \tag{16}$$

Meanwhile, according to the definition of $\theta_\mathcal{S}$ we have

$$R_L(\theta_\mathcal{S}, \boldsymbol{\eta}_\mathcal{S}) < R_L(u, \boldsymbol{\eta}_\mathcal{S}), \tag{17}$$

where $\boldsymbol{\eta}_\mathcal{S}$ is defined in the beginning of the proof, satisfying $\theta_\mathcal{S} \in \Theta^{\boldsymbol{\eta}_\mathcal{S}}$. Denote by $\boldsymbol{\eta}_\lambda \coloneqq (1 - \lambda)\boldsymbol{\eta} + \lambda\boldsymbol{\eta}_\mathcal{S}$. We have $\gamma_\ell(\boldsymbol{\eta}_\lambda) = \mathcal{S}$ for any $\lambda \in [0, 1]$. It follows from $\varphi(u) \notin \mathcal{S}$ and the calibration of L that for every $\lambda \in [0, 1]$ there exists $\theta' \in \mathcal{C}$ such that $R_L(\theta', \boldsymbol{\eta}_\lambda) < R_L(u, \boldsymbol{\eta}_\lambda)$. Meanwhile it follows from the convexity that for any $\lambda \in [0, 1]$ and $\theta', \theta'' \in \mathcal{C}$, we have

$$R_L(\theta', \boldsymbol{\eta}_\lambda) < R_L(u, \boldsymbol{\eta}_\lambda) \;\text{ and }\; R_L(\theta'', \boldsymbol{\eta}_\lambda) < R_L(u, \boldsymbol{\eta}_\lambda) \implies (u - \theta')(u - \theta'') > 0. \tag{18}$$

Combining (17) and (18), we have that $R_L(\theta', \boldsymbol{\eta}_\mathcal{S}) < R_L(u, \boldsymbol{\eta}_\mathcal{S}) \Rightarrow \theta' < u$. Combining (16), (18), and the monotonicity over $\mathcal{C} \cap (-\infty, b]$, we have that $R_L(\theta', \boldsymbol{\eta}) < R_L(u, \boldsymbol{\eta}) \Rightarrow \theta' > u$. According the convexity of L, we have $g_1 < 0 < g_2$ for any $g_1 \in \partial_{\theta=u} R_L(\theta, \boldsymbol{\eta})$ and $g_2 \in \partial_{\theta=u} R_L(\theta, \boldsymbol{\eta}_\mathcal{S})$.

[4] This construction is valid since the intersection on the r.h.s. is non-empty: the calibration allows us to pick some $\theta_0 \in \Theta^{\boldsymbol{\eta}_\mathcal{S}}$. Choose $\theta_1 \in \operatorname{int}(\mathcal{C}) \setminus \{\theta_0\}$ and set $\theta_\epsilon = (1 - \epsilon)\theta_0 + \epsilon\theta_1$. For $0 < \epsilon < 1$, θ_ϵ lies strictly between θ_0 and θ_1, both of which belongs to the interval $\mathcal{C}$, implying $\theta_\epsilon \in \operatorname{int}(\mathcal{C})$; moreover, convexity of L gives $R_L(\theta_\epsilon, \boldsymbol{\eta}_\mathcal{S}) \leq (1 - \epsilon)R_L(\theta_0, \boldsymbol{\eta}_\mathcal{S}) + \epsilon R_L(\theta_1, \boldsymbol{\eta}_\mathcal{S}) < B(\boldsymbol{\eta}_\mathcal{S})$ for sufficiently small $\epsilon > 0$, and thus $\theta_\epsilon \in \Theta^{\boldsymbol{\eta}_\mathcal{S}} \cap \operatorname{int}(\mathcal{C})$.

[5] We have $\boldsymbol{\eta} \neq \boldsymbol{\eta}_{\gamma_\ell(\boldsymbol{\eta})}$ since $L_{|[a,b]}$ is calibrated on $\boldsymbol{\eta}_{\gamma_\ell(\boldsymbol{\eta})}$ according to the definition in the beginning of this proof.

Then suppose $\lambda = \frac{-g_1}{g_2 - g_1} \in [0, 1]$. Note that $(1-\lambda)g_1 + \lambda g_2 \in \partial_{\theta=u} R_L(\theta, \boldsymbol{\eta}_\lambda)$ since R_L is linear in $\boldsymbol{\eta}$, and we have the following inequalities for any $\theta \in \mathcal{C}$ based on the convexity of L that

$$\begin{aligned} R_L(\theta, \boldsymbol{\eta}_\lambda) &\geq R_L(u, \boldsymbol{\eta}_\lambda) + ((1-\lambda)g_1 + \lambda g_2)(\theta - u) \\ &= R_L(u, \boldsymbol{\eta}_\lambda) + \left(\frac{g_2}{g_2 - g_1} g_1 - \frac{g_1}{g_2 - g_1} g_2 \right) (\theta - u) \\ &= R_L(u, \boldsymbol{\eta}_\lambda). \end{aligned}$$

As a result, u minimizes the surrogate risk at $\boldsymbol{\eta}_\lambda$, which indicates that $\varphi(u) \in \operatorname{prop}[L](\boldsymbol{\eta}_\lambda)$. It follows from the calibration of L that $\operatorname{prop}[L](\boldsymbol{\eta}_\lambda) \subseteq \gamma_\ell(\boldsymbol{\eta}_\lambda) = \mathcal{S}$ and thus $\varphi(u) \in \mathcal{S}$, which contradicts with (16). □

G.2 Proof of Theorem 32

Proof. When $\mathcal{C}$ is a singleton, the proof is trivial by extending $L_{|[a,b]}$ as a constant function.

Then we focus on the case that $\mathcal{C}$ is neither empty nor a singleton. According to Lemma 33, we have that there exists $a < b$ and $[a, b] \subseteq \mathcal{C}$ that $(L_{|[a,b]}, \varphi)$ is calibrated with ℓ. It follows from that the closed bounded $[a, b] \subseteq \operatorname{int}(\mathcal{C})$ and Rockafellar [1997, Theorem 10.4] that there exists finite $G > 0$ such that L is G-Lipschitz on $[a, b]$ for any y.

Define

$$L'(\theta, y) \coloneqq \begin{cases} L(a, y) + G(a - \theta), & \theta < a, \\ L(\theta, y), & \theta \in [a, b], \\ L(b, y) + G(\theta - b), & \theta > b, \end{cases} \qquad \varphi'(\theta) = \varphi\big(\Pi_{[a,b]}(\theta)\big),$$

where $\Pi_{[a,b]}$ is the projection onto $[a, b]$. Since L is convex and G-Lipschitz, L' is finite and convex on $\mathbb{R}$.

Then we prove the calibration of (L', φ'). For any $\boldsymbol{\eta} \in \Delta^K$:

$$R_{L'}(\theta, \boldsymbol{\eta}) = R_L(\Pi_P(\theta), \boldsymbol{\eta}) + G|\theta - \Pi_P(\theta)|.$$

Then we have

$$\begin{aligned} \inf_{\theta \in \mathbb{R},\ \varphi'(\theta) \notin \gamma_\ell(\boldsymbol{\eta})} R_{L'}(\theta, \boldsymbol{\eta}) &= \inf_{\theta \in [a,b],\ \varphi(\theta) \notin \gamma_\ell(\boldsymbol{\eta})} R_L(\theta, \boldsymbol{\eta}) \\ &> \min_{\theta \in [a,b]} R_L(\theta, \boldsymbol{\eta}) = \inf_{\theta \in \mathbb{R}} R_{L'}(\theta, \boldsymbol{\eta}), \end{aligned}$$

which indicates the calibration of (L', φ'). □

H Empirical Illustration

We validate our proposed 2-dimensional loss (5) on CIFAR-10 [Krizhevsky, 2009] with rejection costs $c \in \{0.05, 0.10, 0.15, 0.20\}$. We use WideResNet-28-4 [Zagoruyko and Komodakis, 2016] with a two-dimensional output layer and set $\lambda = 5$ and $\tau = 0.25$. The model is trained for 200 epochs using Adam [Kingma and Ba, 2015] for 5 runs, with a learning rate of 10^{-3}, weight decay of 10^{-4}, and batch size of 128. All experiments are conducted on 1 GeForce RTX 4090 GPU. We report the averaged $01c$ risk and acceptance rate in Table 1, where the acceptance rate is the proportion of samples that are not rejected.

Table 1: Results of the proposed 2-dimensional loss on CIFAR-10 with different rejection costs.

Metric	$c = 0.05$	$c = 0.10$	$c = 0.15$	$c = 0.20$
$01c$ risk	0.027	0.041	0.049	0.052
Acceptance rate	0.72	0.85	0.89	0.95

With increasing cost c, the acceptance threshold $1 - c$ in Chow's rule decreases, and our method accepts a larger proportion of samples. Meanwhile, the $01c$ risk increases only moderately, which indicates that the proposed loss can effectively adapt to different rejection costs and make appropriate rejection and classification decisions.

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: The abstract, introduction, and illustration in Figure 1 reflect our scope (design of calibrated surrogate losses) and contributions (a minimal-dimensional surrogate for CwR with $c \in (0, 0.5)$ and a 4-dimensional loss that improve existing loss dimension when $K > 16$.)

   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: The limitation of this work on the potential lower-dimensional surrogates when $c = 0.5$ is discussed as a future direction in the conclusion section.

   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: The assumptions and conditions are stated in the corresponding conclusions, and the proofs are completely provided 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: All the information needed to reproduce the experimental results are included in Appendix H.

   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: [No]

Justification: Our experiments use public datasets and the detailed formulations of the loss is provided, making the work fully reproducible without releasing extra code or data.

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 use the standard train/test split of CIFAR-10, and the optimizer details are provided in Appendix H.

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: [No]

Justification: We report the average results over five independent runs.

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: [Yes]

   Justification: The computer resources used in the experiments are reported in Appendix H.

   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: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   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: [N/A]

    Justification: This is a pure theoretical work focus on the lower bound of loss dimensionality, and we do not identify any specific risks that warrant emphasis.

    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: The paper poses no such risks.

    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: The paper does not use existing assets.

    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: The paper does not release new assets.

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: The paper does not involve crowdsourcing nor research with human subjects.

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: The paper does not involve crowdsourcing nor research with human subjects.

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: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

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.