Establishing Linear Surrogate Regret Bounds for Convex Smooth Losses via Convolutional Fenchel–Young Losses

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

[1] College of Computing and Data Science, Nanyang Technological University
[2] The Institute of Statistical Mathematics
[3] School of Computer Science and Engineering, Southeast University
[4] Skywork AI
yuzhou002@e.ntu.edu.sg bao.han@ism.ac.jp
fenglei@seu.edu.cn boan@ntu.edu.sg

Abstract

Surrogate regret bounds, also known as excess risk bounds, bridge the gap between the convergence rates of surrogate and target losses. The regret transfer is lossless if the surrogate regret bound is linear. While convex smooth surrogate losses are appealing in particular due to the efficient estimation and optimization, the existence of a trade-off between the loss smoothness and linear regret bound has been believed in the community. Under this scenario, the better optimization and estimation properties of convex smooth surrogate losses may inevitably deteriorate after undergoing the regret transfer onto a target loss. We overcome this dilemma for arbitrary discrete target losses by constructing a convex smooth surrogate loss, which entails a linear surrogate regret bound composed with a tailored prediction link. The construction is based on Fenchel–Young losses generated by the *convolutional negentropy*, which are equivalent to the infimal convolution of a generalized negentropy and the target Bayes risk. Consequently, the infimal convolution enables us to derive a smooth loss while maintaining the surrogate regret bound linear. We additionally benefit from the infimal convolution to have a consistent estimator of the underlying class probability. Our results are overall a novel demonstration of how convex analysis penetrates into optimization and statistical efficiency in risk minimization.

1 Introduction

The risk of a machine learning model is often measured by the expectation of a target loss ℓ that quantifies the error between the model prediction t and a natural label y. However, minimizing the target risk over a dataset is often computationally hard because a target prediction problem is usually discretely structured, including multiclass, multilabel, top-k prediction problems. For this reason, a tractable surrogate risk induced by a *surrogate loss* $L(\boldsymbol{\theta}, y)$ serves as an essential proxy with a score $\boldsymbol{\theta} \in \mathbb{R}^d$. The resulting surrogate optimization is no longer discretely constrained.

An ideal surrogate loss should be convex, smooth (or entailing a Lipschitz continuous gradient), and calibrated toward a given target prediction problem. Convexity and smoothness have been fundamental both in optimization and statistical estimation—indeed, classical optimization theory reveals that convex smooth functions can be optimized with first-order methods more efficiently than

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

non-smooth functions [67, 83]. In addition, several studies have demonstrated that the convexity and smoothness can further enhance fast rates in the risk estimation [87, 98]. Meanwhile, calibration is regarded as a minimal requirement on surrogate losses, ensuring that the surrogate risk minimization leads to the target risk minimization [12, 88].[2] To establish a relationship between surrogate and target risks, *surrogate regret bounds* play a crucial role. Therein the regret (or the suboptimality) of a target loss is controlled by that of a surrogate loss through a non-decreasing rate function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. A surrogate regret bound can be informally written as follows: for any score vector $\boldsymbol{\theta} \in \mathbb{R}^d$ and a class distribution $\boldsymbol{\eta}$,

$$\texttt{Regret}_\ell(\varphi(\boldsymbol{\theta}), \boldsymbol{\eta}) \leq \psi\left(\texttt{Regret}_L(\boldsymbol{\theta}, \boldsymbol{\eta})\right), \tag{1}$$

where φ is a prediction link that converts a score vector $\boldsymbol{\theta}$ into a target prediction $\varphi(\boldsymbol{\theta})$. If the regret rate function is linear $\psi(r) = \mathcal{O}(r)$, which is the best possible (data-independent) regret rate,[3] then the optimization and estimation errors of the surrogate loss are optimally translated to the target loss. For example, under the binary classification target loss, Bartlett et al. [12] demonstrated that the linear regret rate is possible with the hinge loss. Later, symmetric losses [24] and polyhedral losses [39] were shown to yield the linear regret rate. Unfortunately, these loss functions lack either convexity or smoothness, which deteriorates the optimization and estimation errors of the surrogate regret, even if they enjoy linear regret bounds. By contrast, a square-root regret rate $\psi(r) = \mathcal{O}(\sqrt{r})$ is common for convex smooth surrogate losses, such as the logistic, exponential, and squared losses [12, 69, 39]. These losses typically enjoy better optimization and estimation properties, yet suffer from larger target regrets due to the suboptimal regret rate. Therefore, it has been open to develop a convex smooth surrogate loss without sacrificing the linear regret rate.

Notwithstanding, the previous literature in this line has implied that such an ideal surrogate loss may be inconceivable. Mahdavi et al. [50] considered this question with an interpolated loss between the hinge (non-smooth) and logistic (smooth) losses and showed that we inevitably face the trade-off between generalization and optimization unless strong distributional assumptions are imposed. Further, Frongillo and Waggoner [39] proved that locally smooth and strongly convex losses must suffer from a square-root regret rate at least. Ramaswamy et al. [76] blame convex smooth losses for redundantly modeling continuous class distributions, which is unnecessary if our goal is merely to solve discrete target problems.

In this paper, we demonstrate that this seemingly impossible trade-off can be overcome for *arbitrary* discrete target losses. Specifically, we build a convex smooth surrogate loss built upon the framework of Fenchel–Young losses [16]—a framework to generate a loss function from a generalized negentropy and its conjugate. In a nutshell, our main results are summarized as follows:

Theorem 1 (Informal version of Theorem 15) *For any discrete target loss ℓ, there exist a convex smooth surrogate loss L (defined over a score $\boldsymbol{\theta} \in \mathbb{R}^d$) and prediction link φ such that the surrogate regret bound* (1) *holds with some linear rate $\psi(r) = \mathcal{O}(r)$.*

This existence is proved constructively, which provides a systematic framework on the construction of convex smooth surrogate losses and their corresponding prediction links with linear surrogate regret bounds. Our high-level construction, which significantly leverages convex analysis and is detailed in Section 3.1, proceeds as follows. First, a user chooses a strongly convex negentropy with some regularity conditions, such as the Shannon negentropy. We encode the structure of a target loss ℓ into the chosen base negentropy by adding the negative Bayes risk of ℓ. This additivity eventually translates into the infimal convolution in the induced Fenchel–Young loss, which we call the *convolutional* Fenchel–Young loss. The convolutional Fenchel–Young loss is endowed with the convexity and smoothness arising from the base negentropy, shown in Section 3.2. Then we can obtain a prediction link via the infimal convolution. Paired with the convolutional Fenchel–Young loss, it admits a linear surrogate regret bound, demonstrated in Section 3.3.

In addition, we improve the multiplicative term in the initial linear surrogate regret bound in Section 3.4 to make the bound tighter, which exploits the low-rank structure of a target loss ℓ. As a by-product of the loss smoothness, we can provide a Fisher-consistent probability estimator of the underlying probability in Section 3.5. Finally, Section 4 instantiates the framework of the convolutional Fenchel–Young loss for the multiclass classification problem, highlighting the efficient computation

[2] Readers should distinguish this calibration property from calibrated prediction [38].

[3] A super-linear bound is possible with distribution-dependent losses [97], which we do not consider here.

of the prediction link. More examples of target prediction problems, such as classification with rejection and multilabel ranking, are available in Appendix D.

1.1 Related Works

Surrogate losses for general discrete prediction problems. A discrete prediction problem aims to predict t that minimizes a discrete target loss $\ell(t, y)$ over the class distribution, and the study of convex surrogates for ℓ is of significant interest. Extensive research has been conducted on surrogate losses for specific discrete tasks, including but not limited to classification [12, 101, 89, 93, 85, 66, 64, 81], top-k [95, 90], and multilabel learning [40, 100, 54, 49, 65].

In contrast to *ad-hoc* approaches, recent studies have advanced the principled design of calibrated convex surrogate losses for general discrete prediction problems, without imposing restrictions on the target losses. In Ramaswamy and Agarwal [73] and Ramaswamy et al. [75], surrogate losses based on the squared loss and error correcting output codes are proposed, respectively, which embed the structure of a target problem into a surrogate loss and are shown to be calibrated with the corresponding target losses. The design of calibrated surrogate losses has also been extensively studied in the context of structured prediction [29, 74, 26, 27, 70, 68, 69, 23], where the structural/low-rank properties of target losses are exploited to construct more efficient surrogates. As opposed to the smooth losses utilized in the works above, polyhedral losses [34, 35, 37] have attracted attention recently, which provides a systematic framework for efficiently constructing calibrated yet non-smooth convex losses based on a discrete target loss.

Surrogate regret bounds. Surrogate regret bounds have been well studied for margin losses under binary and multiclass classification [101, 12, 89, 84, 56, 86], where the target loss is the 0-1 loss. Similarly, proper losses [21, 42, 78, 3, 45, 63] have been shown to provide surrogate regret bounds *w.r.t.* the L_1 distance between the estimated and true class probabilities [77, 8, 9], which facilitates analyses for downstream tasks. For structured prediction, surrogate regret bounds are sometimes called comparison inequalities [26, 70, 68, 14], which are typically of square-root type. Notably, Frongillo and Waggoner [39] demonstrated that polyhedral losses exhibit linear surrogate regret bounds for general discrete prediction problems, covering various piecewise linear non-smooth losses [95, 76] as special cases. However, they lack the smoothness. To the contrary, Mao et al. [53] and Mao et al. [52] propose smooth losses with linear regret rates but lacking convexity.

Whereas a growing line of research has focused on $\mathcal{H}$-consistency to analyze how the restriction to a hypothesis space $\mathcal{H}$ affects consistency [48, 99, 7], we implicitly suppose that the hypothesis space is all measurable functions because the analysis is often more transparent and it is reasonable to suppose that the hypothesis space is sufficiently expressive under the overparametrization regime. The extension to $\mathcal{H}$-consistency is rather straightforward by integrating the minimizability gap [52, 55, 54, 57, 51].[4]

2 Preliminaries

Let $[d] \coloneqq \{1, 2, \cdots, d\}$ and $[\![\cdot]\!]$ is the Iverson bracket. The p-norm is denoted by $\|\cdot\|_p$, which we assume to be the 2-norm unless otherwise noted. Let $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\infty\}$ be the extended real-line and $\Delta^d \coloneqq \{\boldsymbol{\eta} \in \mathbb{R}^d_{\geq 0} : \|\boldsymbol{\eta}\|_1 = 1\}$ the d-simplex. For a set $\mathcal{S} \subseteq \mathbb{R}^d$, $\operatorname{int}(\mathcal{S})$ and $\operatorname{relint}(\mathcal{S})$ are its interior and relative interior, respectively, and $\operatorname{conv}(\mathcal{S})$ is its convex hull. The indicator function is denoted by $\mathbb{I}_{\mathcal{S}} : \mathbb{R}^d \to \{0, +\infty\}$, where $\mathbb{I}_{\mathcal{S}}(\boldsymbol{\theta}) = 0$ if $\boldsymbol{\theta} \in \mathcal{S}$ and $+\infty$ otherwise. For a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$, $\operatorname{dom}(f) \coloneqq \{\boldsymbol{\theta} \in \mathbb{R}^d : f(\boldsymbol{\theta}) < +\infty\}$ is its effective domain. A function f is extended to be set-valued with slight abuse of notation by $f(\mathcal{S}) \coloneqq \{f(s) : s \in \mathcal{S}\}$. The Fenchel conjugate of Ω is $\Omega^*(\boldsymbol{\theta}) \coloneqq \sup_{\boldsymbol{p} \in \operatorname{dom}(\Omega)} \{\boldsymbol{\theta}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})\}$. The identity matrix is I. The canonical basis of the Euclidean space is denoted by $\{\boldsymbol{e}_i\}$, where the dimensionality depends on the contexts.

2.1 Discrete Prediction Problems and Target Losses

Let $\mathcal{Y} = [K]$ be the finite class space. The class distribution on $\mathcal{Y}$ is $\boldsymbol{\eta} \in \Delta^K$ such that class $Y = y$ has probability η_y. A discrete prediction problem aims to find a target prediction t from the finite

[4] This condition can be relaxed to the *well-specified hypothesis space*, under which the hypothesis space is required to contain at least one population risk minimizer, rather than the entire space of measurable functions.

prediction space $\widehat{\mathcal{Y}} \coloneqq [N]$ for each $\boldsymbol{\eta}$ by minimizing a discrete *target loss* $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ over $y \sim \boldsymbol{\eta}$. The averaged target loss is called the target risk: $R_\ell(t, \boldsymbol{\eta}) \coloneqq \mathbb{E}_{y\sim\boldsymbol{\eta}}[\ell(t, y)] = \langle \boldsymbol{\eta}, \boldsymbol{\ell}(t)\rangle$, where $\boldsymbol{\ell}(t) \in \mathbb{R}^K$ is the loss vector such that $\boldsymbol{\ell}(t)_y = \ell(t, y)$ for each $y \in \mathcal{Y}$. The Bayes risk of ℓ at $\boldsymbol{\eta}$ is $\underline{R}_\ell(\boldsymbol{\eta}) \coloneqq \min_{t\in\widehat{\mathcal{Y}}} R_\ell(t, \boldsymbol{\eta})$, which is a concave function of $\boldsymbol{\eta}$. The suboptimality of a prediction t *w.r.t.* the target loss ℓ over a class distribution $\boldsymbol{\eta}$ is characterized by the *target regret*, which is the gap between its risk and the Bayes risk.

Definition 2 (Target regret) *Given a discrete target loss ℓ, the target regret of a prediction t w.r.t. a class probability $\boldsymbol{\eta} \in \Delta^K$ is defined as follows:*

$$\texttt{Regret}_\ell(t, \boldsymbol{\eta}) \coloneqq R_\ell(t, \boldsymbol{\eta}) - \underline{R}_\ell(\boldsymbol{\eta}). \tag{2}$$

Equivalent lower-dimensional decomposition. By encoding every possible class label $y \in \mathcal{Y}$ into the canonical basis $\boldsymbol{e}_y \in \mathbb{R}^K$, we can express any target loss in the following form:

$$\ell(t, y) = \langle \boldsymbol{e}_y, \boldsymbol{\ell}(t)\rangle. \tag{3}$$

An equivalent but more efficient decomposition of (3), known as Structure Encoding Loss Functions (SELF), was introduced in Ciliberto et al. [26], which allows lower-dimensional label encodings. It has been widely used to construct efficient loss functions, and further generalized via Affine Decomposition in Blondel [14, (12)]. We also adopt a general form to represent discrete target losses.

Definition 3 ($(\boldsymbol{\rho}, \boldsymbol{\ell}^{\boldsymbol{\rho}})$-decomposition) *For a discrete target loss $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, its $(\boldsymbol{\rho}, \boldsymbol{\ell}^{\boldsymbol{\rho}})$-decomposition is given as follows:*

$$\ell(t, y) = \langle \boldsymbol{\rho}(y), \boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\rangle + c(y), \tag{4}$$

where $\boldsymbol{\rho} : \mathcal{Y} \to \mathbb{R}^d$ is a label encoding function that maps discrete labels into the d-dimensional Euclidean space, $\boldsymbol{\ell}^{\boldsymbol{\rho}} : \widehat{\mathcal{Y}} \to \mathbb{R}^d$ is the corresponding loss encoding function, and $c : \mathcal{Y} \to \mathbb{R}$ is the remainder independent of prediction t.

This loss decomposition contains the Affine Decomposition [14, (12)]. Any discrete target loss admits such a decomposition via the trivial choice $\boldsymbol{\rho}(y) = \boldsymbol{e}_y$, $\boldsymbol{\ell}^{\boldsymbol{\rho}}(t) = \boldsymbol{\ell}(t)$, and $c = 0$ with $d = K$, which immediately recovers (3). With a properly chosen $\boldsymbol{\rho}$, the encoded cardinality d can often be greatly smaller than K, particularly in the context of structured prediction. For example, consider multilabel classification, where $\mathcal{Y} = \widehat{\mathcal{Y}} = [K]$, $K = 2^d$, and each $y \in [K]$ corresponds to a unique binary vector $\boldsymbol{\nu}(y) \in \{0, 1\}^d$ representing a set of binary labels. Suppose our target loss is the Hamming loss $\ell_H(t, y) \coloneqq \mathbf{1}^\top(\boldsymbol{\nu}(t) + \boldsymbol{\nu}(y)) - 2\langle \boldsymbol{\nu}(t), \boldsymbol{\nu}(y)\rangle$, which is the Hamming distance between $\boldsymbol{\nu}(t)$ and $\boldsymbol{\nu}(y)$. This entails the decomposition $\boldsymbol{\rho}(y) = \mathbf{1} - 2\boldsymbol{\nu}(y)$, $\boldsymbol{\ell}^{\boldsymbol{\rho}}(t) = \boldsymbol{\nu}(t)$, and $c(y) = \mathbf{1}^\top\boldsymbol{\nu}(y)$. Here, $d = \log_2 K$ significantly reduces the dimensionality from the cardinality of $\mathcal{Y}$. We refer reader to Blondel [14, Appendix A] and Appendix D for more examples of discrete prediction problems. The cardinality d of label encoding $\boldsymbol{\rho}$ is closely related to surrogate regret bounds later in Section 3.4.

2.2 Surrogate Losses and Surrogate Regret Bounds

Unlike discrete target losses assessing a discrete prediction $t \in \widehat{\mathcal{Y}}$, a surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \overline{\mathbb{R}}$ receives a continuous score $\boldsymbol{\theta} \in \mathbb{R}^d$. Then a prediction link $\varphi : \mathbb{R}^d \to \widehat{\mathcal{Y}}$ is used to transform a score $\boldsymbol{\theta}$ to a prediction t. Its risk and Bayes risk are defined as $R_L(\boldsymbol{\theta}, \boldsymbol{\eta}) \coloneqq \mathbb{E}_{y\sim\boldsymbol{\eta}}[L(\boldsymbol{\theta}, y)] = \langle \boldsymbol{\eta}, \boldsymbol{L}(\boldsymbol{\theta})\rangle$ and $\underline{R}_L(\boldsymbol{\eta}) \coloneqq \inf_{\boldsymbol{\theta}\in\mathbb{R}^d} R_L(\boldsymbol{\theta}, \boldsymbol{\eta})$, respectively, where $\boldsymbol{L}(\boldsymbol{\theta}) \coloneqq [L(\boldsymbol{\theta}, y)]_{y=1}^K$. The regret of a surrogate loss is defined as the gap between the risk of a score $\boldsymbol{\theta}$ and the Bayes risk, similar to the target regret.

Definition 4 (Surrogate regret) *Given a surrogate loss L, the surrogate regret of score $\boldsymbol{\theta}$ w.r.t. a class probability $\boldsymbol{\eta} \in \Delta^K$ is defined as follows:*

$$\texttt{Regret}_L(\boldsymbol{\theta}, \boldsymbol{\eta}) \coloneqq R_L(\boldsymbol{\theta}, \boldsymbol{\eta}) - \underline{R}_L(\boldsymbol{\eta}). \tag{5}$$

For a desirable surrogate loss, the convergence of its surrogate regret will dominate the target regret. That is, for a fixed class distribution $\boldsymbol{\eta}$, a convergent $\boldsymbol{\theta}$ such that $\texttt{Regret}_L(\boldsymbol{\theta}, \boldsymbol{\eta}) \to 0$ should imply $\texttt{Regret}_\ell(\varphi(\boldsymbol{\theta}), \boldsymbol{\eta}) \to 0$. This convergence relationship indicates that the target regret minimization can be achieved by the minimization of the surrogate regret adopted with an appropriate prediction link. A surrogate regret bound offers a quantitative characterization of this relationship through a regret rate function ψ, which is the key focus of this work.

Definition 5 (Surrogate regret bound) *A surrogate loss L and prediction link φ entail a surrogate regret bound w.r.t. target ℓ with a non-decreasing regret rate function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfying $\psi(0) = 0$ if the following inequality holds:*

$$\texttt{Regret}_\ell(\varphi(\boldsymbol{\theta}), \boldsymbol{\eta}) \leq \psi(\texttt{Regret}_L(\boldsymbol{\theta}, \boldsymbol{\eta})), \quad \text{for any } (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^d \times \Delta^K. \tag{6}$$

While a linear regret rate $\psi(r) = \mathcal{O}(r)$ is the best possible, previously know linear-rate losses are either non-smooth or non-convex, including the (non-smooth) hinge loss [12] and (non-convex) sigmoid loss [24]. These loss functions may face challenges in optimization and estimation [87]. In this work, we aim to develop a framework that facilitates the use of convex smooth surrogates with linear surrogate regret bounds.

2.3 Fenchel–Young Loss

In this work, we build upon Fenchel–Young losses [16, 18, 59] to construct convex smooth surrogates equipped with linear regret rate functions. Let us review Fenchel–Young losses first.

Definition 6 (Fenchel–Young loss) *For $\Omega : \mathbb{R}^d \to \overline{\mathbb{R}}$, the associated Fenchel–Young loss $L_\Omega : \mathrm{dom}(\Omega^*) \times \mathrm{dom}(\Omega) \to \mathbb{R}_{\geq 0}$ is defined as follows:*

$$L_\Omega(\boldsymbol{\theta}, \boldsymbol{p}) = \Omega(\boldsymbol{p}) + \Omega^*(\boldsymbol{\theta}) - \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle. \tag{7}$$

Similar constructions can also be found in Duchi et al. [33], Agarwal et al. [2], which focus on multiclass classification. The function Ω is often regarded as a generalized negentropy, which equals the negative of the Bayes risk of its induced Fenchel–Young loss. Fenchel–Young losses often impose additional requirements on the domains of Ω and Ω^*, tailored to specific target problems, and we will highlight them when necessary.

Fenchel–Young losses possess favorable properties: they are inherently convex (in score $\boldsymbol{\theta}$) by definition. Moreover, we can use the map $\nabla\Omega^*(\boldsymbol{\theta})$ to obtain the mean $\mathbb{E}_{y\sim\boldsymbol{\eta}}[\boldsymbol{\rho}(y)]$, which is the class distribution $\boldsymbol{\eta}$ under multiclass classification with $\boldsymbol{\rho}(y) = \boldsymbol{e}_y$; or linear properties in general [1]. Many common losses are encompassed in Fenchel–Young losses, including the cross-entropy loss, squared loss, and Crammer–Singer loss [31]. For example, $\Omega(\boldsymbol{p}) = \frac{1}{2}\|\boldsymbol{p}\|^2 + \mathbb{I}_{\Delta^K}(\boldsymbol{p})$ generates the sparsemax loss, which induces the sparsemax as a Fisher-consistent estimator of $\boldsymbol{\eta}$ [58].

Connections to information geometry. While Fenchel–Young losses are systematically defined under the convex-analytic formulation, it also has a longstanding history in information geometry, particularly through its connection to the Bregman divergence and related generalizations [72]. In Blondel et al. [16], it is noted that a Fenchel–Young loss can be interpreted as the mixed-type Bregman divergence [4]. When Ω is of Legendre-type, Ω can yield the dual coordinate system and the Fenchel–Young loss is further equivalent to the *canonical divergence* generated by Ω [5, Eq. (3.44)].

3 Convex Smooth Surrogates with Linear Surrogate Regret Bounds

In this section, we first show our construction of surrogate losses and prediction links. After showing that this loss is convex and smooth with an appropriately chosen base negentropy (Theorem 11), we move on to the main result of this paper, linear surrogate regret bounds with convex smooth surrogate losses (Theorem 13). We further discuss the computational aspects and improve the regret bound constant. The missing proofs are deferred to Appendix B.

3.1 Construction

We design a convex smooth surrogate loss built upon the framework of Fenchel–Young losses. To make its surrogate regret bound linear, we craft a negentropy by delicately leveraging the structure of a target prediction problem. Denote by T the polyhedral convex function $T(\boldsymbol{p}) = -\min_{t\in\widehat{\mathcal{Y}}} \langle \boldsymbol{p}, \boldsymbol{\ell}^{\boldsymbol{\rho}}(t) \rangle$. Then, we have the following definition.

Definition 7 (Convolutional negentropy) *Suppose $(\boldsymbol{\rho}, \boldsymbol{\ell}^{\boldsymbol{\rho}})$-decomposition for a target loss ℓ. For a negentropy $\Omega : \mathbb{R}^d \to \overline{\mathbb{R}}$, its convolutional negentropy is defined as follows:*

$$\Omega_T(\boldsymbol{p}) = \Omega(\boldsymbol{p}) + T(\boldsymbol{p}) \tag{8}$$

While commonly used negentropy in Fenchel–Young losses, such as the Shannon negentropy and norm negentropy [19], are unstructured toward a target loss, the convolutional negentropy Ω_T encodes a target loss ℓ explicitly into the base negentropy Ω via T. This polyhedral convex function T is an affinely transformed negative Bayes risk of a target loss ℓ when $\boldsymbol{p} \in \operatorname{conv}\{\boldsymbol{\rho}(\mathcal{Y})\}$. Indeed, with the linear property $\boldsymbol{p} = \mathbb{E}_{y\sim\boldsymbol{\eta}}[\boldsymbol{\rho}(y)]$, i.e., expectation of $\boldsymbol{\rho}(y)$ *w.r.t.* class probability $\boldsymbol{\eta}$:

$$T(\boldsymbol{p}) = -\min_{t\in\widehat{\mathcal{Y}}} \langle \boldsymbol{p}, \boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\rangle = -\min_{t\in\widehat{\mathcal{Y}}} \mathbb{E}_{y\sim\boldsymbol{\eta}}[\langle \boldsymbol{\rho}(y), \boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\rangle] = \sum_{y\in[K]} \eta_y c(y) - \underline{R}_\ell(\boldsymbol{\eta}). \tag{9}$$

The base negentropy Ω is up to our choice. Before deriving the conjugate of Ω_T, we make the following requirement on Ω throughout this work for a well-behaved convolutional negentropy.

Condition 1 *Ω is proper convex and lower-semicontinuous (l.s.c.), and satisfies* $\operatorname{conv}(\boldsymbol{\rho}(\mathcal{Y})) \subseteq \operatorname{dom}(\Omega)$ *and* $\operatorname{dom}(\Omega^*) = \mathbb{R}^d$.

It is a mild requirement. Indeed, proper convexity and lower-semicontinuity merely aim to avoid pathologies, and the domain assumptions are met by a differentiable and finite Ω over the valid prediction space $\operatorname{conv}(\boldsymbol{\rho}(\mathcal{Y}))$. The assumption on Ω^* is crucial for the induced loss to have unconstrained domain $\mathbb{R}^d$. Then we show an explicit form of the conjugated convolutional negentropy.

Lemma 8 (Conjugate of Ω_T) *Suppose Condition 1 holds, then Ω_T^* is proper convex and l.s.c. with $\operatorname{dom}(\Omega_T^*) = \mathbb{R}^d$, and can be expressed as follows:*

$$\Omega_T^*(\boldsymbol{\theta}) = \inf_{\boldsymbol{\pi}\in\Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) \quad \textit{for any } \boldsymbol{\theta} \in \mathbb{R}^d, \tag{10}$$

where $\mathcal{L}^{\boldsymbol{\rho}} \in \mathbb{R}^{d\times N}$ is the loss matrix with the t-th column being $\boldsymbol{\ell}^{\boldsymbol{\rho}}(t) \in \mathbb{R}^d$ for $t \in \widehat{\mathcal{Y}}$. In addition, there always exists $\boldsymbol{\pi} \in \Delta^N$ that achieves the infimum of (10) *for any $\boldsymbol{\theta} \in \mathbb{R}^d$.*

The conjugated form (10) owes to the *infimal convolution* between the base negentropy Ω and the target Bayes risk T. We coined the name of the convolutional negentropy inspired by this structure. Intuitively, the conjugated convolutional negentropy Ω_T^* can be viewed as a "perturbed" version of the conjugated base negentropy Ω^* toward the direction of the target loss matrix $\mathcal{L}^{\boldsymbol{\rho}}$. This result facilitates a more concrete formulation of the Fenchel–Young loss generated by Ω_T, which we refer to as the convolutional Fenchel–Young loss—the central focus of this work.

Definition 9 (Convolutional Fenchel–Young loss) *Suppose that a discrete target loss ℓ enjoys $(\boldsymbol{\rho}, \boldsymbol{\ell}^{\boldsymbol{\rho}})$-decomposition. For a negentropy $\Omega : \mathbb{R}^d \to \overline{\mathbb{R}}$ satisfying Condition 1, the convolutional Fenchel–Young loss $L_{\Omega_T} : \mathbb{R}^d \times \mathcal{Y} \to \overline{\mathbb{R}}$ induced by the convolutional negentropy Ω_T is defined as follows:*

$$L_{\Omega_T}(\boldsymbol{\theta}, y) = \min_{\boldsymbol{\pi}\in\Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) + \Omega_T(\boldsymbol{\rho}(y)) - \langle \boldsymbol{\theta}, \boldsymbol{\rho}(y)\rangle. \tag{11}$$

Note that $L_{\Omega_T}(\boldsymbol{\theta}, \cdot)$ is deliberately constrained to the class space $\mathcal{Y}$ instead of the general domain $\operatorname{dom}(\Omega_T)$, to align with the surrogate loss form introduced in Section 2.2.

Given a surrogate loss, we need to specify a prediction link. For general discrete prediction problems where $\mathcal{Y} \neq \widehat{\mathcal{Y}}$, a prediction link yields a discrete prediction in $\widehat{\mathcal{Y}}$ from a score $\boldsymbol{\theta} \in \mathbb{R}^d$ obtained through minimizing the surrogate loss. Thus, a loss and link are the two sides of the same coin. Unlike the standard argmax link in multiclass classification [89, 94], we create an alternative argmax-like prediction link based on the minimizer of (10).

Definition 10 ($\boldsymbol{\pi}$-argmax link) *Let $\Pi : \mathbb{R}^d \to 2^{\Delta^N}$ be a set-valued map defined as follows:*

$$\Pi(\boldsymbol{\theta}) \coloneqq \operatorname{argmin}\left\{\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) : \boldsymbol{\pi} \in \Delta^N\right\} \quad \textit{for any } \boldsymbol{\theta} \in \mathbb{R}^d. \tag{12}$$

Let $\boldsymbol{\pi} : \mathbb{R}^d \to \Delta^N$ be a selector of Π such that $\boldsymbol{\pi}(\boldsymbol{\theta}) \in \Pi(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$.[5] *Then the $\boldsymbol{\pi}$-argmax link $\varphi : \mathbb{R}^d \to \widehat{\mathcal{Y}}$ is defined as*

$$\varphi(\boldsymbol{\theta}) \in \operatorname{argmax}\left\{\pi_t(\boldsymbol{\theta}) : t \in \widehat{\mathcal{Y}}\right\} \quad \textit{for any } \boldsymbol{\theta} \in \mathbb{R}^d,$$

where the tie can be broken arbitrarily.

[5] We slightly abuse the notation $\boldsymbol{\pi}$ by using it for both a vector and a vector-valued mapping; in particular, the $\boldsymbol{\pi}$ appearing in the $\boldsymbol{\pi}$-argmax link refers to the selector function.

The existence of such links is guaranteed as long as $\Pi(\boldsymbol{\theta})$ is non-empty for every $\boldsymbol{\theta} \in \mathbb{R}^d$, which is guaranteed by Lemma 8. While the standard argmax link returns a prediction $t \in \widehat{\mathcal{Y}}$ with the maximum score θ_t, the $\boldsymbol{\pi}$-argmax link returns t with the maximum π_t. This probabilistic quantity $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$ defined via (10) distorts the original score $\boldsymbol{\theta}$ by leveraging the target loss structure $\mathcal{L}^{\boldsymbol{\rho}}$.

3.2 Convexity and Smoothness

Before discussing surrogate regret bounds, we verify that convolutional Fenchel–Young losses are indeed convex and smooth, implied by the conjugacy between smoothness and strong convexity [44].

Corollary 11 (Convexity and smoothness of L_{Ω_T}) *Suppose that a discrete target loss ℓ enjoys $(\boldsymbol{\rho}, \boldsymbol{\ell}^{\boldsymbol{\rho}})$-decomposition. For a base negentropy Ω, we additionally suppose that Condition 1 is satisfied. If Ω is strictly convex on $\operatorname{dom}(\Omega)$, $L_{\Omega_T}(\cdot, y)$ is convex and differentiable over $\mathbb{R}^d$ for any $y \in \mathcal{Y}$. If Ω is additionally strongly convex on $\operatorname{dom}(\Omega)$, $L_{\Omega_T}(\cdot, y)$ is smooth over $\mathbb{R}^d$ for any $y \in \mathcal{Y}$.*

There are several exemplar base negentropies fulfilling the conditions of Corollary 11. For example, the squared norm $\Omega(\boldsymbol{p}) = \frac{1}{2}\|\boldsymbol{p}\|^2$ is strongly convex with the self-conjugate $\Omega^*(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2$. In this case, Ω is effective over the entire $\mathbb{R}^d$ and hence includes $\operatorname{conv}(\boldsymbol{\rho}(\mathcal{Y}))$, satisfying Condition 1. In light of the target-loss decomposition (3), we have $\operatorname{conv}(\boldsymbol{\rho}(\mathcal{Y})) = \Delta^K$. In this case, the Shannon negentropy $\Omega(\boldsymbol{p}) = \langle \boldsymbol{p}, \ln \boldsymbol{p} \rangle + \mathbb{I}_{\Delta^K}(\boldsymbol{p})$ is strongly convex on Δ^K with its conjugate $\Omega^*(\boldsymbol{\theta}) = \ln\langle \exp(\boldsymbol{\theta}), \mathbf{1} \rangle$ satisfying $\operatorname{dom}(\Omega^*) = \mathbb{R}^K$, where $\ln$ and $\exp$ are element-wise. Hence, both base negentropy yield convex and smooth L_{Ω_T}.

Note that L_{Ω_T} is not locally strongly convex at every point. We can see this by noting that T is not differentiable and neither is Ω_T, which implies that Ω_T^* is not strictly convex [80, Theorem 11.13]. This is important for establishing linear surrogate regret bounds (in Section 3.3) because the known square-root regret rate lower bound considers locally strongly convex surrogate losses [39, Theorem 2].

While the loss L_{Ω_T} is now ensured to be convex and smooth, its gradient calculation, which is the basis for gradient-based optimization [17, 15], remains non-trivial. This is because the gradient of $\min_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi})$ contained in (11) cannot be written analytically in general. We show that its gradient calculation reduces to computing $\Pi(\boldsymbol{\theta})$ through the following variant of envelope theorems.

Lemma 12 (Envelope theorem) *Suppose Condition 1 holds and Ω is strictly convex on $\operatorname{dom}(\Omega)$. For any $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$, we have*

$$\nabla_{\boldsymbol{\theta}} \Big[\min_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) \Big] = \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}). \tag{13}$$

Deviating from the standard envelope theorems [15, 13], we carefully addresses the non-uniqueness of the minimizers. Given this, the gradient of the convolutional Fenchel–Young loss (11) is accessed via $\nabla_{\boldsymbol{\theta}} L_{\Omega_T}(\boldsymbol{\theta}, y) = \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) - \boldsymbol{\rho}(y)$. At the core of its proof, the differentiability of Ω_T^* (indirectly obtained via Theorem 11) is vital.

3.3 Linear Surrogate Regret Bounds

Now we exhibit linear surrogate regret bounds with the convolutional Fenchel–Young loss.

Theorem 13 (Linear surrogate regret bound) *Consider a target loss with $(\boldsymbol{\rho}, \boldsymbol{\ell}^{\boldsymbol{\rho}})$-decomposition. For a negentropy $\Omega : \mathbb{R}^d \to \overline{\mathbb{R}}$, suppose Condition 1 holds. For any $\boldsymbol{\pi}$-argmax link φ, (L_{Ω_T}, φ) admits the following surrogate regret bound:*

$$\texttt{Regret}_{\ell}(\varphi(\boldsymbol{\theta}), \boldsymbol{\eta}) \leq N \texttt{Regret}_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}), \quad \textit{for any } (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^d \times \Delta^K. \tag{14}$$

The regret bound in Theorem 13 indeed has the linear rate $\psi(r) = Nr$ in the form (1), while L_{Ω_T} can be naturally made convex and smooth as discussed in Section 3.2. This is a remarkable consequence because many authors have previously implied the impossibility to overcome the barrier of the square-root regret rate with convex smooth surrogate losses [50, 76, 39, 8]. The crux of this success lies in the infimal convolution structure in (10), which enables us to additively decompose the conjugate Ω_T^*, exploited below. The constant N will be improved in Section 3.4.

From now on we sketch the regret bound proof. The following lemma is a cornerstone herein.

Lemma 14 (Lower bound of surrogate regret) *Assume the same set of conditions as in Theorem 13. For any $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$, we have the following inequality:*

$$\sum_{t=1}^{N} \pi_t \texttt{Regret}_\ell(t, \boldsymbol{\eta}) \le \texttt{Regret}_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}), \quad \textit{for any } (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^d \times \Delta^K. \tag{15}$$

Its proof, formally given in Appendix B, hinges on the following additive regret decomposition:

$$\texttt{Regret}_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \underbrace{R_{L_\Omega}(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}, \boldsymbol{\eta})}_{\text{Risk of Fenchel–Young loss } L_\Omega \ge 0} + \underbrace{\sum\nolimits_{t \in \widehat{\mathcal{Y}}} \pi_t \texttt{Regret}_\ell(t, \boldsymbol{\eta})}_{\text{Convex combination of the target regret}} \quad \forall (\boldsymbol{\theta}, \boldsymbol{\pi}) \in \mathbb{R}^d \times \Pi(\boldsymbol{\theta}),$$

which directly implies the lower bound (15) because the Fenchel–Young loss L_Ω is non-negative. Note that this *additive* decomposition is indispensable for the linear lower bound (15). This becomes possible just because we create the convolutional negentropy Ω_T based on the additive form in (8), which yields the additive form $\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}$ in the conjugate expression (10). All of these are thanks to the structure of the infimal convolution, and Theorem 13 immediately follows with rescaling.

Finally, we achieve the main result by combining Corollary 11 and Theorem 13. We constructively prove the existence of convex smooth linear-regret surrogate losses, by noting that both strongly convex Ω satisfying Condition 1 and $\boldsymbol{\pi}$-argmax link do exist.

Theorem 15 (Main result) *Consider a target loss ℓ with $(\boldsymbol{\rho}, \boldsymbol{\ell}^{\boldsymbol{\rho}})$-decomposition. For a negentropy $\Omega : \mathbb{R}^d \to \overline{\mathbb{R}}$ satisfying Condition 1, suppose that Ω is strongly convex on $\operatorname{dom}(\Omega)$. Then the convolutional Fenchel–Young loss $L_{\Omega_T}(\cdot, y)$ is convex, differentiable, and smooth over $\mathbb{R}^d$ for any $y \in \mathcal{Y}$. Moreover, with $\boldsymbol{\pi}$-argmax link φ, (L_{Ω_T}, φ) enjoys the linear surrogate regret bound* (14)*.*

3.4 Improving Constant of Linear Surrogate Regret Bounds

The constant N in the linear surrogate regret bound in (14) can be prohibitively large, leading to a vacuous bound. This issue is significant in structured prediction as studied in Osokin et al. [70]. For example, in multilabel classification with the Hamming loss, $N = 2^d$ is the number of all the potential binary label predictions that increases exponentially with d, the number of binary labels. In top-k classification, $N = \binom{K}{k}$ is the number of all possible size-k subsets of the class space. Thankfully, the geometry property of the problem (10) provides a promising scheme for reducing the dependency on prediction number N, which induces the following improved link.

Corollary 16 (Improved surrogate regret bound) *Suppose Condition 1 holds. There exists $\tilde{\boldsymbol{\pi}} : \mathbb{R}^d \to \Delta^N$ such that $\tilde{\boldsymbol{\pi}}(\boldsymbol{\theta}) \in \Pi(\boldsymbol{\theta})$ and $\|\tilde{\boldsymbol{\pi}}(\boldsymbol{\theta})\|_0 \le \operatorname{affdim}(\mathcal{L}^{\boldsymbol{\rho}}) + 1$ for any $\boldsymbol{\theta} \in \mathbb{R}^d$. Moreover, with the induced $\tilde{\boldsymbol{\pi}}$-argmax link $\varphi_{\tilde{\boldsymbol{\pi}}} : \mathbb{R}^d \to \Delta^N$, $(L_{\Omega_T}, \varphi_{\tilde{\boldsymbol{\pi}}})$ admits the following surrogate regret bound:*

$$\texttt{Regret}_\ell(\varphi_{\tilde{\boldsymbol{\pi}}}(\boldsymbol{\theta}), \boldsymbol{\eta}) \le [\operatorname{affdim}(\mathcal{L}^{\boldsymbol{\rho}}) + 1]\, \texttt{Regret}_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}), \quad \textit{for any } (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^d \times \Delta^K, \tag{16}$$

where $\operatorname{affdim}(\mathcal{L}^{\boldsymbol{\rho}})$ is the dimension of the affine hull of the column vectors of $\mathcal{L}^{\boldsymbol{\rho}}$.

Note that we have $\operatorname{affdim}(\mathcal{L}^{\boldsymbol{\rho}}) \le \min\{N, d\}$, which indicates that the dependency on N can be largely reduced when $d \ll N$. For example, in top-k classification, $\boldsymbol{\rho}(y) = \boldsymbol{e}_y$ is used with $d = K$, which is much smaller than $N = \binom{K}{k}$. In multilabel classification with Hamming loss, $d = \log_2 N$ is logarithmically smaller than N when binary encoding $\boldsymbol{\rho}(y) = \boldsymbol{\nu}(y)$ is used (see Section 2.1).

3.5 Bonus: Fisher-consistent Probability Estimator

We discuss a benefit of convex smooth surrogate losses beyond discrete prediction problems. Oftentimes people are interested in a probability estimator over possible prediction outcomes, recovering from surrogate risk minimizers, as in classification with rejection [25, 11, 28, 30, 22, 55, 36]. Herein non-smooth surrogate losses have been unfavored because of lacking reasonable probability estimators [60, 24]. Ramaswamy et al. [76] conjectures the incompatibility of probability estimation with linear regret bounds. By contrast, we give a Fisher-consistent estimator of the linear property $\mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)]$ for convolutional Fenchel–Young losses without sacrificing the linear regret bound.

Theorem 17 (Fisher-consistent probability estimator) *Suppose Condition 1 holds and Ω is strictly convex on $\operatorname{dom}(\Omega)$. For any $\boldsymbol{\eta} \in \operatorname{relint}(\Delta^K)$, the surrogate risk $R_{L_{\Omega_T}}(\cdot, \boldsymbol{\eta})$ is minimized at $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that $\nabla\Omega^*(\boldsymbol{\theta}^* + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) = \mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)]$ for any $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta}^*)$.*

Algorithm 1 Exact Solution of (17) in $\mathcal{O}(K \ln K)$

Sort $\boldsymbol{\theta} \in \mathbb{R}^K$ such that $\theta_{(1)} \geq \cdots \geq \theta_{(K)}$.
$n \leftarrow \max\{k \in [K] : \ 1 + k\theta_{(k)} > \sum_{i=1}^{k} \theta_{(i)}\}$.
$\tau(\boldsymbol{\theta}) \leftarrow \frac{\sum_{i=1}^{n} \theta_{(i)} - 1}{n}$.
$\pi_{\log}(\boldsymbol{\theta})_i \leftarrow \max\{\theta_i - \tau(\boldsymbol{\theta}), 0\}$.

As a result, the empirical minimizers w.r.t. L_{Ω_T} are Fisher-consistent estimators of the linear property $\mathbb{E}_{y\sim\boldsymbol{\eta}}[\boldsymbol{\rho}(y)]$. According to the result above, we can first solve $\Pi(\boldsymbol{\theta})$ in (12) and select $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$. Then, $\nabla\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi})$ is a rational estimate of the linear property $\mathbb{E}_{y\sim\boldsymbol{\eta}}[\boldsymbol{\rho}(y)]$. When we follow the trivial loss decomposition (3) with $\boldsymbol{\rho}(y) = \boldsymbol{e}_y$, the estimand is $\mathbb{E}_{y\sim\boldsymbol{\eta}}[\boldsymbol{\rho}(y)] = \boldsymbol{\eta} \in \Delta^K$. For this reason, we say $\nabla\Omega^*(\boldsymbol{\theta}^* + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi})$ is a "probability" estimator.

For example, let us recap the encoding of multilabel classification in Section 2.1, where $y \in \mathcal{Y}$ is a multilabel among all 2^d possible combinations of d binary labels. A multilabel y is encoded by $\boldsymbol{\nu}(y) \in \{0,1\}^d$, and the Hamming loss is decomposed with $\boldsymbol{\rho}(y) = 1 - 2\boldsymbol{\nu}(y)$. Here $\mathbb{E}_{y\sim\boldsymbol{\eta}}[\boldsymbol{\nu}(y)]$ reads $\bar{\boldsymbol{\nu}} = [\mathrm{Prob}(\boldsymbol{\nu}_i(y) = 1)]_{i=1}^d$, whose i-th element indicates the likelihood of the binary class i being positive. Through the relation $\boldsymbol{\rho} = 1 - 2\boldsymbol{\nu}$, the probability estimator $\nabla\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi})$ is capable of recovering $\bar{\boldsymbol{\nu}}$ by $[1 - \nabla\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi})]/2$.

4 Example: Multiclass Classification

We demonstrate convolutional Fenchel–Young losses for multiclass classification. and further examples of prediction problems can be found in Appendices D and E, including detailed computational complexity analyses and visualizations of the associated binary classification losses. In multiclass classification, the class space and the prediction space are the same: $\mathcal{Y} = \widehat{\mathcal{Y}} = [K]$. For an input with class probability $\boldsymbol{\eta} \in \Delta^K$, the goal is to predict the most likely class $t \in \mathrm{argmax}_{t\in\mathcal{Y}} \eta_t$. The target loss is the 0-1 loss $\ell_{01}(t, y) = [\![t \neq y]\!]$.

Firstly, we adopt the decomposition (3) for this task, that is, $\boldsymbol{\ell}^{\boldsymbol{\rho}}(t) = [\ell_{01}(t,1), \cdots, \ell_{01}(t,K)] = \mathbf{1} - \boldsymbol{e}_t$ and $\boldsymbol{\rho}(y) = \boldsymbol{e}_y$, which corresponds to the one-hot encoding commonly used in this task. In this case, $N = d = K$, and $\mathcal{L}^{\boldsymbol{\rho}} = \mathbf{1}\mathbf{1}^\top - I$.

Next, we move on to the convolutional negentropy Ω_T. For the choice of Ω, we use the Shannon negentropy, which induces the celebrated cross-entropy loss in the original Fenchel–Young loss framework [16, Table 1]. Its conjugate is the log-sum-exp function $\Omega^*(\boldsymbol{\theta}) = \ln\langle\exp(\boldsymbol{\theta}), \mathbf{1}\rangle$. Then we can calculate the conjugate of Ω_T based on Lemma 8:

$$\Omega_T^*(\boldsymbol{\theta}) = \min_{\boldsymbol{\pi}\in\Delta^K} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) = \min_{\boldsymbol{\pi}\in\Delta^K} \ln\langle\exp(\boldsymbol{\theta} + \mathbf{1} - \boldsymbol{\pi}), \mathbf{1}\rangle\,. \tag{17}$$

Since Ω is proper convex and l.s.c. with $\mathrm{conv}(\boldsymbol{\rho}(\mathcal{Y})) = \Delta^K = \mathrm{dom}(\Omega)$ and $\mathrm{dom}(\Omega^*) = \mathbb{R}^K$, Ω satisfies Condition 1. Furthermore, since Ω is strongly convex on Δ^K, we can finally derive the convolutional Fenchel–Young loss (11) by nothing $\Omega_T(\boldsymbol{\rho}(y)) = 0$ for all $y \in \mathcal{Y}$, as follows:

$$L_{\Omega_T}(\boldsymbol{\theta}, y) = \Omega_T^*(\boldsymbol{\theta}) + \Omega_T(\boldsymbol{\rho}(y)) - \langle\boldsymbol{\theta}, \boldsymbol{\rho}(y)\rangle = \min_{\boldsymbol{\pi}\in\Delta^K} \ln\langle\exp(\boldsymbol{\theta} + \mathbf{1} - \boldsymbol{\pi}), \mathbf{1}\rangle - \theta_y, \tag{18}$$

which is convex and smooth in $\boldsymbol{\theta}$ thanks to Theorem 11. While shares the form of cross-entropy loss $L_\Omega(\boldsymbol{\theta}, y) = \ln\langle\exp(\boldsymbol{\theta}), \mathbf{1}\rangle - \theta_y$, it further incorporates an additional bounded perturbation.

Solving the minimization problem (17) is important from the computational aspects, including the gradient calculation of L_{Ω_T} and accessing to the probability estimator given by Theorem 17. We provide Algorithm 1 to solve (17) with $\mathcal{O}(K \ln K)$ time.

Lemma 18 *For any $\boldsymbol{\theta} \in \mathbb{R}^K$, the problem* (17) *has a unique minimizer $\boldsymbol{\pi}_{\log}(\boldsymbol{\theta}) \in \Pi(\boldsymbol{\theta})$, which can be obtained in $\mathcal{O}(K \ln K)$ time by Algorithm 1.*

The proof is deferred to Appendix C.1, following from the KKT conditions. Eventually we have

$$\nabla\Omega_T^*(\boldsymbol{\theta}) = \mathrm{softmax}\left(\boldsymbol{\theta} + \mathbf{1} - \boldsymbol{\pi}_{\log}(\boldsymbol{\theta})\right), \ \text{ where } \ \mathrm{softmax}(\boldsymbol{\theta})_y = \frac{\exp(\theta_y)}{\sum_{i=1}^K \exp(\theta_i)} \ \text{ for } y \in [K].$$

Algorithm 1 comes with a significant resemblance with the sparsemax [58, Algorithm 1], which is determined by the similar structure shared by (17) and Euclidean projection problem [58, (2)]. Since $\Pi(\boldsymbol{\theta})$ is a singleton, we have the unique $\boldsymbol{\pi}_{\log}$-argmax link $\varphi_{\boldsymbol{\pi}_{\log}}$ (Definition 10). While we need to solve problem (17) for every $\boldsymbol{\theta}$ to have access to $\nabla_{\boldsymbol{\theta}} L_{\Omega_T}(\boldsymbol{\theta}, y)$, the prediction by the $\boldsymbol{\pi}_{\log}$-argmax link is much cheaper, without requiring Algorithm 1, which indicates that we can simply use the class label with the largest score $\max_{y \in [K]} \theta_y$ in test time. The proof can be found in Appendix C.2.

Proposition 19 *A prediction link φ is the $\boldsymbol{\pi}_{\log}$-argmax link if and only if $\varphi(\boldsymbol{\theta}) \in \operatorname{argmax}_{t \in \mathcal{Y}} \theta_t$.*

Remark 20 *Under classification, a surrogate loss with a regret bound is classification-calibrated, which indicates that the surrogate risk minimization eventually leads to the Bayes-optimal classifier. In literature, Blondel [14] and Wang and Scott [94] have investigated sufficient conditions for Fenchel–Young losses to be classification-calibrated. Therein the base negentropy is assumed to be of Legendre-type or twice differentiable. Our convolutional Fenchel–Young losses are interesting because we do not require these conditions to yield both the smoothness and a regret bound. Thus it remains open to relax these existing sufficient conditions for Fenchel–Young losses further.*

5 Discussion

Convex non-smooth linear regret losses. Compared to existing convex non-smooth surrogates with linear regret, e.g., polyhedral losses [37], our smooth convex surrogate has several advantages. First, smoothness enables more efficient optimization, as supported by results in both deterministic and stochastic optimization regimes [67, 83] while it is left open to exploit specific structures arising from polyhedral losses to achieve faster optimization. In addition, the smoothness can lead to optimistic ERM rates of estimation error [87], offering better estimation in easier tasks. Another appealing aspect is that our loss admits a consistent probability estimator (Theorem 17), which can be valuable for downstream tasks such as uncertainty quantification and calibration.

Efficient gradient calculation. In general discrete prediction problems, we need to solve the minimization (10) to take a gradient of L_{Ω_T}, which is potentially demanding over a high-dimensional domain Δ^N. To have access to the gradient, we can alternatively solve

$$\min_{\boldsymbol{\nu} \in V} \Omega^*(\boldsymbol{\theta} + \boldsymbol{\nu}), \quad \text{where } V \coloneqq \operatorname{conv}\Big(\{\boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\}_{t \in [N]}\Big). \tag{19}$$

To see this, let us denote the minimizer of (19) by $\boldsymbol{\nu}^*$. Then (19) is an equivalent optimization problem to (10) because there exists $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$ such that $\boldsymbol{\nu}^* = \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}$. Eventually $\nabla \Omega^*(\boldsymbol{\theta} + \boldsymbol{\nu}^*)$ serves as an alternative to the gradient formula (13). Thus we can reduce the dimensionality of the optimization problem. For example, in multilabel classification (Section 2.1), $V = [0, 1]^d$ has a logarithmically smaller optimization dimensionality than $N = 2^d$. We can solve (19) with this box constraint efficiently by using the standard L-BFGS solver [47].

Randomized prediction link. Recall that the $\boldsymbol{\pi}$-argmax link φ (Definition 10) deterministically outputs in the probability simplex Δ^N. Instead we can define a randomized link $\tilde{\varphi}$ such that $\Pr[\tilde{\varphi}(\boldsymbol{\theta}) = t] = \pi_t(\boldsymbol{\theta})$. This yields a better regret bound by Lemma 14, as follows:

$$\mathbb{E}_{\tilde{\varphi}(\boldsymbol{\theta})}[\mathtt{Regret}_{\ell}(\tilde{\varphi}(\boldsymbol{\theta}), \boldsymbol{\eta})] = \sum_{t=1}^{N} \pi_t(\boldsymbol{\theta})\, \mathtt{Regret}_{\ell}(t, \boldsymbol{\eta}) \leq \mathtt{Regret}_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \quad \forall (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^d \times \Delta^K,$$

where the expectation is taken over the randomness of $\tilde{\varphi}$. Although we adopt the *expected* target regret differently from Theorem 13, the constant of regret bounds is strikingly improved from N to 1, which is dimension-free. Sakaue et al. [82] observes a similar regret improvement by a randomized link for online structured prediction with Fenchel–Young losses.

6 Conclusion

In this work, we construct convex and smooth surrogate losses with linear surrogate regret bounds by leveraging Fenchel–Young losses and infimal convolution. Our results demonstrate that convexity, smoothness, and linear surrogate regret are compatible for arbitrary discrete prediction problems. Moreover, our loss naturally admits consistent probability estimator, bridging the gap between linear regret and estimation. We illustrate the broad applicability of our approach through examples in multiclass classification, classification with rejection, and multilabel ranking. Overall, this study highlights the utility of convex analysis as a principled tool for designing surrogate losses.

Acknowledgement

We thank Shinsaku Sakaue for his valuable insights on the target-loss decomposition and for his careful review, and we also thank the anonymous reviewers for their attentive reading of the manuscript and their many thoughtful comments and suggestions. This research is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 Award MOE-T2EP20223-0003. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore. YC is also supported by Google PhD Fellowship program. HB is supported by JST PRESTO (Grant No. JPMJPR24K6), Japan. Lei Feng is supported by the Big Data Computing Center of Southeast University.

References

[1] 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: 5

[2] Alekh Agarwal, Sham Kakade, Nikos Karampatziakis, Le Song, and Gregory Valiant. Least squares revisited: Scalable approaches for multi-class prediction. In *International Conference on Machine Learning*, pages 541–549, 2014. Cited on page: 5

[3] Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(1):1653–1674, 2014. Cited on page: 3

[4] Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194. Springer, 2016. Cited on page: 5

[5] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191. American Mathematical Soc., 2000. Cited on page: 5

[6] Pranjal Awasthi, Natalie Frank, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Calibration and consistency of adversarial surrogate losses. *Advances in Neural Information Processing Systems*, 34:9804–9815, 2021. Cited on page: 25

[7] Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-class H-consistency bounds. *Advances in Neural Information Processing Systems*, 35:782–795, 2022. Cited on page: 3

[8] Han Bao. Proper losses, moduli of convexity, and surrogate regret bounds. In *Conference on Learning Theory*, pages 525–547, 2023. Cited on pages: 3, 7

[9] Han Bao and Asuka Takatsu. Proper losses regret at least 1/2-order. *arXiv preprint arXiv:2407.10417*, 2024. Cited on page: 3

[10] Han Bao, Clayton Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *Conference on Learning Theory*, pages 408–451, 2020. Cited on page: 25

[11] 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: 8

[12] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. Cited on pages: 2, 3, and 5

[13] Dimitri Bertsekas. *Nonlinear Programming*, volume 4. Athena Scientific, 2016. Cited on page: 7

[14] Mathieu Blondel. Structured prediction with projection oracles. *Advances in Neural Information Processing Systems*, 32:12145–12156, 2019. Cited on pages: 3, 4, and 10

[15] Mathieu Blondel and Vincent Roulet. The elements of differentiable programming. *arXiv preprint arXiv:2403.14606*, 2024. Cited on page: 7

[16] 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 pages: 2, 5, and 9

[17] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *Advances in Neural Information Processing Systems*, 35:5230–5242, 2022. Cited on page: 7

[18] Mathieu Blondel, Felipe Llinares-López, Robert Dadashi, Léonard Hussenot, and Matthieu Geist. Learning energy networks with generalized Fenchel–Young losses. *Advances in Neural Information Processing Systems*, 35:12516–12528, 2022. Cited on page: 5

[19] Dick E. Boekee and Jan C. A. van der Lubbe. The R-norm information measure. *Information and Control*, 45(2):136–155, 1980. Cited on page: 6

[20] Jonathan M. Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer New York, 2005. Cited on page: 25

[21] 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: 3

[22] 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: 8

[23] Yuzhou Cao, Lei Feng, and Bo An. Consistent hierarchical classification with a generalized metric. In *International Conference on Artificial Intelligence and Statistics*, pages 4825–4833, 2024. Cited on page: 3

[24] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning*, pages 961–970, 2019. Cited on pages: 2, 5, and 8

[25] Chao-Kong Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. Cited on pages: 8, 30

[26] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. *Advances in Neural Information Processing Systems*, 29:4412–4420, 2016. Cited on pages: 3, 4

[27] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 21(98):1–67, 2020. Cited on page: 3

[28] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82, 2016. Cited on page: 8

[29] Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. *Advances in Neural Information Processing Systems*, 29:2514–2522, 2016. Cited on page: 3

[30] 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 page: 8

[31] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001. Cited on page: 5

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. Cited on page: 37

[33] John Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018. Cited on page: 5

[34] 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: 3

[35] Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. Embedding dimension of polyhedral losses. In *Conference on Learning Theory*, pages 1558–1585, 2020. Cited on page: 3

[36] Jessie 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 page: 8

[37] Jessie 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: 3, 10

[38] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998. Cited on page: 2

[39] Rafael Frongillo and Bo Waggoner. Surrogate regret bounds for polyhedral losses. *Advances in Neural Information Processing Systems*, 35:21569–21580, 2021. Cited on pages: 2, 3, 7, 25, and 37

[40] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Conference on Learning Theory*, pages 341–358, 2011. Cited on page: 3

[41] Aritra Ghosh, Himanshu Kumar, and P. Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial intelligence*, volume 31, 2017. Cited on page: 25

[42] 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: 3

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. Cited on page: 37

[44] Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. Technical report, Toyota Technological Institute at Chicago, 2009. Cited on page: 7

[45] Wojciech Kotłowski and Krzysztof Dembczyński. Surrogate regret bounds for generalized classification performance metrics. In *Asian Conference on Machine Learning*, pages 301–316, 2016. Cited on page: 3

[46] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. Cited on page: 38

[47] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989. Cited on page: 10

[48] Phil Long and Rocco Servedio. Consistency versus realizable H-consistency for multiclass classification. In *International Conference on Machine Learning*, pages 801–809, 2013. Cited on page: 3

[49] Michal Lukasik, Lin Chen, Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, Felix X Yu, Sashank J Reddi, Gang Fu, Mohammadhossein Bateni, and Sanjiv Kumar. Bipartite ranking from multiple labels: On loss versus label aggregation. *arXiv preprint arXiv:2504.11284*, 2025. Cited on page: 3

[50] Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Binary excess risk for smooth convex surrogates. *arXiv preprint arXiv:1402.1792*, 2014. Cited on pages: 2, 7

[51] Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds: Characterization and extensions. *Advances in Neural Information Processing Systems*, 36:4470–4508, 2023. Cited on page: 3

[52] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning*, pages 23803–23828, 2023. Cited on pages: 3, 25

[53] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Structured prediction with stronger consistency guarantees. *Advances in Neural Information Processing Systems*, 36:46903–46937, 2023. Cited on pages: 3, 25

[54] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-label learning with stronger consistency guarantees. *Advances in neural information processing systems*, 37:2378–2406, 2024. Cited on page: 3

[55] 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, 2024. Cited on pages: 3, 8

[56] Anqi Mao, Mehryar Mohri, and Yutao Zhong. A universal growth rate for learning with smooth surrogate losses. *Advances in Neural Information Processing Systems*, 37:41670–41708, 2024. Cited on page: 3

[57] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Enhanced H-consistency bounds. In *International Conference on Algorithmic Learning Theory*, pages 772–813, 2025. Cited on page: 3

[58] André F. T. Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623, 2016. Cited on pages: 5, 10

[59] André F. T. Martins, Marcos Treviso, António Farinhas, Pedro M. Q. Aguiar, Mário A. T. Figueiredo, Mathieu Blondel, and Vlad Niculae. Sparse continuous distributions and Fenchel–Young losses. *Journal of Machine Learning Research*, 23(257):1–74, 2020. Cited on page: 5

[60] Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: Theory, robustness to outliers, and SavageBoost. *Advances in Neural Information Processing Systems*, 21:1049–1056, 2008. Cited on page: 8

[61] Nishant A. Mehta and Robert C. Williamson. From stochastic mixability to fast rates. *Advances in Neural Information Processing Systems*, 27:1197–1205, 2014. Cited on page: 25

[62] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 54(8):3797–3803, 2008. Cited on page: 25

[63] Aditya Krishna Menon and Robert C. Williamson. Bayes-optimal scorers for bipartite ranking. In *Conference on Learning Theory*, pages 68–106, 2014. Cited on page: 3

[64] Aditya Krishna Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pages 603–611, 2013. Cited on page: 3

[65] Aditya Krishna Menon, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Multilabel reductions: What is my loss optimising? *Advances in Neural Information Processing Systems*, 32:10600–10611, 2019. Cited on page: 3

[66] Harikrishna Narasimhan, Harish G. Ramaswamy, Shiv Kumar Tavker, Drona Khurana, Praneeth Netrapalli, and Shivani Agarwal. Consistent multiclass algorithms for complex metrics and constraints. *Journal of Machine Learning Research*, 25(367):1–81, 2024. Cited on page: 3

[67] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl Akad Nauk SSSR*, 269:543, 1983. Cited on pages: 2, 10

[68] Alex Nowak, Francis Bach, and Alessandro Rudi. Sharp analysis of learning with discrete losses. In *International Conference on Artificial Intelligence and Statistics*, pages 1920–1929, 2019. Cited on page: 3

[69] Alex Nowak, Francis Bach, and Alessandro Rudi. A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*, 2019. Cited on pages: 2, 3

[70] Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. *Advances in Neural Information Processing Systems*, 30: 302–313, 2017. Cited on pages: 3, 8

[71] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8024–8035, 2019. Cited on page: 37

[72] Seta Rakotomandimby, Jean-Philippe Chancelier, Michel De Lara, and Mathieu Blondel. Learning with Fitzpatrick losses. *Advances in Neural Information Processing Systems*, 37: 79381–79409, 2024. Cited on page: 5

[73] Harish G. Ramaswamy and Shivani Agarwal. Classification calibration dimension for general multiclass losses. *Advances in Neural Information Processing Systems*, 25:2078–2086, 2012. Cited on pages: 3, 28

[74] Harish G. Ramaswamy, Shivani Agarwal, and Ambuj Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. *Advances in Neural Information Processing Systems*, 26:1475–1483, 2013. Cited on page: 3

[75] Harish G. Ramaswamy, Balaji S. Babu, Shivani Agarwal, and Robert C. Williamson. On the consistency of output code based learning algorithms for multiclass learning problems. In *Conference on Learning Theory*, pages 885–902, 2014. Cited on page: 3

[76] 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: 2, 3, 7, and 8

[77] Mark D. Reid and Robert C. Williamson. Surrogate regret bounds for proper losses. In *International Conference on Machine Learning*, pages 897–904, 2009. Cited on page: 3

[78] Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010. Cited on page: 3

[79] Ralph T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997. Cited on pages: 25, 26, and 28

[80] Ralph T. Rockafellar and Roger J.-B. R. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009. Cited on pages: 7, 26, 27, and 28

[81] Vincent Roulet, Tianlin Liu, Nino Vieillard, Michael E. Sander, and Mathieu Blondel. Loss functions and operators generated by f-divergences. In *International Conference on Machine Learning*, volume 267, pages 52110–52138, 2025. Cited on page: 3

[82] Shinsaku Sakaue, Han Bao, Taira Tsuchiya, and Taihei Oki. Online structured prediction with Fenchel–Young losses and improved surrogate regret for online multiclass classification with logistic loss. In *Conference on Learning Theory*, pages 4458–4486, 2024. Cited on page: 10

[83] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017. Cited on pages: 2, 10

[84] Clayton Scott. Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In *International Conference on Machine Learning*, pages 153–160, 2011. Cited on page: 3

[85] Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6: 958–992, 2012. Cited on page: 3

[86] Sanket Shah, Milind Tambe, and Jessie Finocchiaro. Analyzing cost-sensitive surrogate losses via H-calibration. *arXiv preprint arXiv:2502.19522*, 2025. Cited on page: 3

[87] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in Neural Information Processing Systems*, 24:2199–2207, 2010. Cited on pages: 2, 5, 10, and 25

[88] Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007. Cited on page: 2

[89] Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007. Cited on pages: 3, 6

[90] Anish Thilagar, Rafael Frongillo, Jessica Finocchiaro, and Emma Goodwill. Consistent polyhedral surrogates for top-k classification and variants. In *International Conference on Machine Learning*, pages 21329–21359, 2022. Cited on page: 3

[91] Tim van Erven, Peter D. Grünwald, Mark D. Reid, and Robert C. Williamson. Mixability in statistical learning. *Advances in Neural Information Processing Systems*, 26:1700–1708, 2012. Cited on page: 25

[92] Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015. Cited on page: 25

[93] Yutong Wang and Clayton Scott. On classification-calibration of gamma-phi losses. In *Conference on Learning Theory*, pages 4929–4951, 2023. Cited on page: 3

[94] Yutong Wang and Clayton Scott. Unified binary and multiclass margin-based classification. *Journal of Machine Learning Research*, 25(143):1–51, 2024. Cited on pages: 6, 10

[95] Forest Yang and Sanmi Koyejo. On the consistency of top-k surrogate losses. In *International Conference on Machine Learning*, volume 119, pages 10727–10735, 2020. Cited on page: 3

[96] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016. Cited on page: 38

[97] Jingwei Zhang, Tongliang Liu, and Dacheng Tao. On the rates of convergence from surrogate risk minimizers to the Bayes optimal classifier. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5766–5774, 2021. Cited on page: 2

[98] Lijun Zhang, Tianbao Yang, and Rong Jin. Empirical risk minimization for stochastic convex optimization: $O(1/n)$- and $O(1/n^2)$-type of risk bounds. In *Conference on Learning Theory*, pages 1954–1979, 2017. Cited on page: 2

[99] Mingyuan Zhang and Shivani Agarwal. Bayes consistency vs. H-consistency: The interplay between surrogate loss functions and the scoring function class. *Advances in Neural Information Processing Systems*, 33:16927–16936, 2020. Cited on page: 3

[100] Mingyuan Zhang, Harish G. Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label F-measure. In *International Conference on Machine Learning*, pages 11246–11255, 2020. Cited on page: 3

[101] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004. Cited on page: 3

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 and Theorem 1 in the introduction (Section 1) reflect our scope (surrogate regret bound of losses) and contribution (a family of convex smooth surrogates with linear surrogate regret bounds).

   Guidelines:

   - The answer NA 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 NA 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 is discussed in Appendix A.

   Guidelines:

   - The answer NA 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: Assumptions are clearly stated in each theoretical claim. All theoretical claims are provided with proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The experimental setting, including the data augmentation, optimizer, model architecture, and associated computational cost.

   Guidelines:

   - The answer NA 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, the official PyTorch training script, and detailed formulations of both the loss and solver, making the work fully reproducible without releasing extra code or data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings are provided in Appendix F.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: For ImageNet experiments, we perform a 5% t-test for comparison. For classification with rejection results, no comparison is involved, and we report only the averaged performance.

   Guidelines:

   - The answer NA 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 detailed information on computer resources is provided in Appendix F.

   Guidelines:

   - The answer NA 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:

   Guidelines:

   - The answer NA 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: [NA]

    Justification: This theoretical work aims to advance machine learning by informing the design and use of surrogate losses for discrete prediction. While it may influence downstream research and applications, we do not identify any specific risks that warrant emphasis.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA 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: [NA]

    Justification:

    Guidelines:

    - The answer NA 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: [NA]

    Justification:

    Guidelines:

    - The answer NA 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: [NA]

    Justification:

    Guidelines:

    - The answer NA 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: [NA]

    Justification:

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Additional Discussions on Related Losses

Fast rate losses. Let us briefly discuss the target risk estimation error rates induced by different surrogate losses. Suppose that a surrogate risk can be estimated with the estimation error upper bound ϵ_{est} and the surrogate regret rate is ψ. Then, the target risk estimation error is of order $\psi(\epsilon_{\text{est}})$. Hence, we need to take care of both ψ and ϵ_{est} to discuss the target risk estimation.

On the one hand, loss functions entailing either strongly convex or exponentially concave are known to achieve fast estimation error rates with ERM (e.g., $\mathcal{O}(1/n)$) in standard parametric setting, where the hypothesis class is of finite dimension [87, 91, 61, 92]. However, Frongillo and Waggoner [39, Theorem 4] reveals that typical fast rate losses suffer at least from square-root regret rates, and thus their corresponding target risk estimation error bounds are $\mathcal{O}(1/\sqrt{n})$ at least. On the other hand, convolutional Fenchel–Young losses (which is convex smooth but not strongly convex) yield the estimation error upper bounds of order $\mathcal{O}(1/\sqrt{n})$ under the same setting, while the final target risk upper bounds are in order of $\mathcal{O}(1/\sqrt{n})$ thanks to the linear regret rate function. As a result, convolutional Fenchel–Young losses achieve target regret convergence rates that are comparable to those of existing fast-rate losses.

In more general nonparametric settings, where fast rates are often hard to achieve without additional assumptions [62], convolutional Fenchel–Young losses achieve a target risk estimation error bound of order $\mathcal{O}(1/\sqrt{n})$. In contrast, strongly convex surrogate losses typically achieve a slower target risk estimation error rate than $\mathcal{O}(1/\sqrt{n})$ because the general nonparametric estimation error is $\mathcal{O}(1/\sqrt{n})$ but the regret rate are slower than $\psi(r) = \mathcal{O}(r)$.

While we do not intend to argue that the convolutional Fenchel–Young loss is always better, we would like to highlight that it may be a good alternative when the parametric conditions for fast rates cannot be readily justified. It also remains an open and worthwhile question whether fast rates can be obtained for our loss under additional assumptions, such as low-noise or margin conditions.

Smooth non-convex linear regret losses. While our focus is on convex surrogates due to their favorable optimization and statistical properties, we note that certain smooth but non-convex surrogates, such as the mean absolute error (MAE) [41, 52] and structured comp-sum losses with MAE [53], also achieve linear regret. These methods, while typically non-convex and not equipped with probability estimator, offer valuable advantages in other aspects, such as robustness to label noise, $\mathcal{H}$-consistency, and potential benefits under adversarial conditions [10, 6], where non-convexity can play a meaningful role.

B Deferred Proofs in Section 3

Recalling the definition of $T(\boldsymbol{p}) = -\min_{t\in\widehat{\mathcal{Y}}} \langle \boldsymbol{p}, \boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\rangle$, which is the negative of the affinely transformed target Bayes risk in (9). It can be inferred that it is proper convex and l.s.c., and we have $\Omega_T = \Omega + T$.

B.1 Proof of Lemma 8

Proof. Since Ω and T are proper convex and l.s.c., so are Ω_T and thus Ω_T^*. Then we prove (10) using the infimal convolution. First, by noting that T is nothing else but the support function of the closed convex set $\operatorname{conv}(\{-\boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\}_{t\in\widehat{\mathcal{Y}}})$, we can express T^* as follows:

$$T^*(\boldsymbol{\theta}) = \sup_{\boldsymbol{p}\in\mathbb{R}^d} \left[\langle \boldsymbol{\theta}, \boldsymbol{p}\rangle - \max_{t\in\widehat{\mathcal{Y}}} \langle \boldsymbol{p}, -\boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\rangle\right] = \mathbb{I}_{\operatorname{conv}\left(\{-\boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\}_{t\in\widehat{\mathcal{Y}}}\right)}(\boldsymbol{\theta}),$$

where we use the conjugacy relationship between a support function and indicator function of a closed convex set $\operatorname{conv}(\{-\boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\}_{t\in\widehat{\mathcal{Y}}})$ [79, Section 13]. According to Condition 1 and the definition of T, we see that both Ω and T are proper convex and T is continuous on $\operatorname{dom}(\Omega)$. Then we use the infimal convolution [20] to derive the conjugate of Ω_T:

$$\begin{aligned}\Omega_T^*(\boldsymbol{\theta}) &= \inf_{\boldsymbol{\theta}'\in\mathbb{R}^d} \left[\Omega^*(\boldsymbol{\theta}-\boldsymbol{\theta}') + \mathbb{I}_{\operatorname{conv}\left(\{-\boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\}_{t\in\widehat{\mathcal{Y}}}\right)}(\boldsymbol{\theta}')\right] \\ &= \inf_{\boldsymbol{\theta}'\in\operatorname{conv}\left(\{-\boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\}_{t\in\widehat{\mathcal{Y}}}\right)} \Omega^*(\boldsymbol{\theta}-\boldsymbol{\theta}')\end{aligned}$$

$$= \inf_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}).$$

Next we show that the infimum is indeed achieved by some $\boldsymbol{\pi} \in \Delta^N$. Note that for any $\boldsymbol{\theta} \in \mathbb{R}^d$, the set $\Theta \coloneqq \{\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi} : \boldsymbol{\pi} \in \Delta^N\}$ is compact and non-empty, and that Ω^* is l.s.c. on Θ since $\operatorname{dom}(\Omega^*) = \mathbb{R}^d$. Then the infimum of $\Omega^*(\boldsymbol{\theta})$ is achieved by some $\boldsymbol{\theta}'' \in \Theta$, and there must exist $\boldsymbol{\pi} \in \Delta^N$ such that $\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi} = \boldsymbol{\theta}''$ by the definition of Θ, which complete the proof of the existence of the minimizer.

Finally we see that for any $\boldsymbol{\theta} \in \mathbb{R}^d$, there exists $\boldsymbol{\pi} \in \Delta^N$ such that $\Omega_T^*(\boldsymbol{\theta}) = \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})$, which is smaller than $+\infty$ since $\operatorname{dom}(\Omega^*) = \mathbb{R}^d$ and $\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi} \in \mathbb{R}^d$. This indicates that $\operatorname{dom}(\Omega_T^*) = \mathbb{R}^d$ because $\boldsymbol{\theta}$ is chosen arbitrarily. ■

B.2 Proof of Theorem 11

Proof. In the convolutional Fenchel–Young loss (11), the term $\Omega_T(\boldsymbol{\rho}(y)) - \langle \boldsymbol{\theta}, \boldsymbol{\rho}(y) \rangle$ is linear. Hence it suffices to prove the convexity and smoothness of the conjugate $\min_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})$.

When Ω is strictly convex, Ω_T is also strictly convex since T is convex. Since the strict convexity holds on $\operatorname{dom}(\Omega_T)$, Ω_T is further essentially strictly convex, that is, strictly convex on every convex subset of $\operatorname{dom}(\partial \Omega_T)$ [79, p253]. According to Rockafellar [79, Theorem 26.3], Ω_T^* is essentially smooth, that is, differentiable throughout non-empty $\operatorname{int}(\operatorname{dom}(\Omega_T^*))$ with $\|\nabla \Omega_T^*(\boldsymbol{\theta})\|$ diverging to $+\infty$ when $\boldsymbol{\theta}$ approaches a boundary point of $\operatorname{int}(\operatorname{dom}(\Omega_T^*))$, which immediately indicates the differentiability on $\mathbb{R}^d$.

When Ω is further strongly convex, so is Ω_T. According to Condition 1 and Lemma 8, Ω_T is proper convex and l.s.c., and thus the biconjugate Ω_T^{**} matches Ω_T by the Fenchel–Moreau theorem. According to Rockafellar and Wets [80, Proposition 12.60], Ω_T^* is smooth since its conjugate $\Omega_T^{**} = \Omega_T$ is strongly convex. ■

B.3 Proof of Lemma 12

Proof. According to Condition 1, the strict convexity of Ω, and the proof of Theorem 11, both Ω_T^* and Ω^* are differentiable on $\mathbb{R}^d$. In addition, we have $\operatorname{dom}(\Omega) = \operatorname{dom}(\Omega_T)$ according to (8).

Note that $\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})$ is convex in $\boldsymbol{\pi}$. Then its first-order optimal condition for any $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$ reads

$$\left\langle \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}), \mathcal{L}^{\boldsymbol{\rho}}(\boldsymbol{\pi}' - \boldsymbol{\pi}) \right\rangle \geq 0 \qquad \text{any } \boldsymbol{\pi}' \in \Delta^N.$$

First, for any $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$, we choose

$$t \in \operatorname{argmin}_{t' \in \widehat{\mathcal{Y}}} \langle \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}), \boldsymbol{\ell}^{\boldsymbol{\rho}}(t') \rangle,$$

then we have

$$\begin{aligned}
\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) &\overset{\text{(A)}}{=} (\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})^\top \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) - \Omega(\nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})) \\
&= \boldsymbol{\theta}^\top \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) + \langle \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}), \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi} \rangle - \Omega(\nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})) \\
&\overset{\text{(A)}}{\leq} \boldsymbol{\theta}^\top \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) + \langle \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}), \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{e}_t \rangle - \Omega(\nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})) \\
&= \boldsymbol{\theta}^\top \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) + \langle \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}), \boldsymbol{\ell}^{\boldsymbol{\rho}}(t) \rangle - \Omega(\nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})) \\
&\overset{\text{(B)}}{=} \boldsymbol{\theta}^\top \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) + \underbrace{\min_{t' \in \widehat{\mathcal{Y}}} \langle \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}), \boldsymbol{\ell}^{\boldsymbol{\rho}}(t') \rangle}_{= -T(\nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}))} - \Omega(\nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})) \\
&\overset{\text{(C)}}{=} \boldsymbol{\theta}^\top \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) - \Omega_T(\nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi})) \\
&\overset{\text{(D)}}{\leq} \sup_{\boldsymbol{p} \in \operatorname{dom}(\Omega_T)} \boldsymbol{\theta}^\top \boldsymbol{p} - \Omega_T(\boldsymbol{p}) \\
&= \Omega_T^*(\boldsymbol{\theta}),
\end{aligned}$$

where (A) owes to the optimality of $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$, (B) holds by the definition of t, and (C) holds by the definition of the convolutional negentropy Ω_T (8). Since $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$, we have $\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) = \Omega_T^*(\boldsymbol{\theta})$ according to Lemma 8. Thus the above inequality is indeed an identity. In particular, (D) becomes an identity, which implies that the supremum of $\sup_{\boldsymbol{p} \in \mathrm{dom}(\Omega_T)} \boldsymbol{\theta}^\top \boldsymbol{p} - \Omega_T(\boldsymbol{p})$ is achieved at $\nabla\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi})$. Since the supremum is also achieved at $\nabla\Omega_T^*(\boldsymbol{\theta})$ and the maximizer is unique according to Rockafellar and Wets [80, Proposition 11.3] and the differentiability of Ω_T^*, we have

$$\nabla\Omega_T^*(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}).$$

Since $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \in \mathbb{R}^d$ are chosen arbitrarily, this concludes the proof. ■

B.4 Proof of Theorem 13

Proof. Fix any $\boldsymbol{\theta} \in \mathbb{R}^d$ and choose any $\boldsymbol{\pi} \in \Delta^N$ out of $\Pi(\boldsymbol{\theta})$. Then we have

$$\begin{aligned}
\forall \boldsymbol{\eta} \in \Delta^K,\ \ \mathtt{Regret}_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) &\geq \sum\nolimits_{t=1}^N \pi_t \mathtt{Regret}_\ell(t, \boldsymbol{\eta}) && \text{(by (15))} \\
&\geq \pi_{\varphi(\boldsymbol{\theta})} \mathtt{Regret}_\ell(\varphi(\boldsymbol{\theta}), \boldsymbol{\eta}) && \text{(because } \pi_t \geq 0 \text{ for any } t \in \widehat{\mathcal{Y}}\text{)} \\
&\geq \mathtt{Regret}_\ell(\varphi(\boldsymbol{\theta}), \boldsymbol{\eta})/N. && \text{(by definition of } \varphi \text{ in Defintion 10)}
\end{aligned}$$

■

B.5 Proof of Lemma 14

Proof. First, we derive the Bayes risk of the convolutional Fenchel–Young loss L_{Ω_T} as follows:

$$\begin{aligned}
\underline{R}_{L_{\Omega_T}}(\boldsymbol{\eta}) &= \inf_{\boldsymbol{\theta} \in \mathbb{R}^d} R_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \\
&= \inf_{\boldsymbol{\theta} \in \mathbb{R}^d} \Big[\Omega_T^*(\boldsymbol{\theta}) - \langle \boldsymbol{\theta}, \mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)] \rangle\Big] + \mathbb{E}_{y \sim \boldsymbol{\eta}}[\Omega_T(\boldsymbol{\rho}(y))] \\
&= -\sup_{\boldsymbol{\theta} \in \mathbb{R}^d} \Big[\langle \boldsymbol{\theta}, \mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)] \rangle - \Omega_T^*(\boldsymbol{\theta})\Big] + \mathbb{E}_{y \sim \boldsymbol{\eta}}[\Omega_T(\boldsymbol{\rho}(y))] \\
&= -\Omega_T^{**}\Big(\mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)]\Big) + \mathbb{E}_{y \sim \boldsymbol{\eta}}[\Omega_T(\boldsymbol{\rho}(y))] \\
&= \mathbb{E}_{y \sim \boldsymbol{\eta}}[\Omega_T(\boldsymbol{\rho}(y))] - \Omega_T\Big(\mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)]\Big)
\end{aligned}$$

where the Fenchel–Moreau theorem is applied to proper convex and l.s.c. Ω_T^* (by Lemma 8) at the last identity. Then we can get the following regret lower bound for any $\boldsymbol{\pi} \in \Pi(\boldsymbol{\theta})$:

$$\begin{aligned}
\mathtt{Regret}_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) &= R_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \underline{R}_{L_{\Omega_T}}(\boldsymbol{\eta}) \\
&= \Omega_T^*(\boldsymbol{\theta}) - \Big\langle \boldsymbol{\theta}, \mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)] \Big\rangle + \Omega_T\Big(\mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)]\Big) \\
&= \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) - \Big\langle \boldsymbol{\theta}, \mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)] \Big\rangle + \Omega\Big(\mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)]\Big) + T\Big(\mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)]\Big) \\
&\overset{\text{(A)}}{=} \underbrace{\Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) - \Big\langle \boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}, \mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)] \Big\rangle + \Omega\Big(\mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)]\Big)}_{\geq 0 \quad \text{due to the Fenchel–Young inequality}} \\
&\quad + \Big\langle \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}, \mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)] \Big\rangle + T\Big(\mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)]\Big) \\
&\overset{\text{(B)}}{\geq} \sum_{t=1}^N \pi_t \Big\langle \mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)], \boldsymbol{\ell}^{\boldsymbol{\rho}}(t) \Big\rangle - \min_{t \in \widehat{\mathcal{Y}}} \Big\langle \mathbb{E}_{y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(y)], \boldsymbol{\ell}^{\boldsymbol{\rho}}(t) \Big\rangle \\
&\overset{\text{(C)}}{=} \sum_{t=1}^N \pi_t \mathtt{Regret}_\ell(t, \boldsymbol{\eta}),
\end{aligned}$$

where (A) holds according to the explicit form of Ω_T^* in (10) and the definition of $\Pi(\boldsymbol{\theta})$, (B) owes to Fenchel–Young inequality, and (C) holds according to the definition of $(\boldsymbol{\rho}, \boldsymbol{\ell}^{\boldsymbol{\rho}})$-decomposition in (4). This concludes the proof. ■

B.6 Proof of Corollary 16

Proof. Fix an arbitrary $\boldsymbol{\theta} \in \mathbb{R}^d$. Since $\Pi(\boldsymbol{\theta})$ is non-empty, we can select $\boldsymbol{\pi}$ from $\Pi(\boldsymbol{\theta})$. Note that $\mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi} \in \operatorname{conv}(\{\boldsymbol{\ell}^{\boldsymbol{\rho}}(t)\}_{t\in\widehat{\mathcal{Y}}})$. According to Carathéodory's theorem, there exists $\tilde{\boldsymbol{\pi}}$ such that $\|\tilde{\boldsymbol{\pi}}\|_0 \leq \operatorname{affdim}(\mathcal{L}^{\boldsymbol{\rho}}) + 1$ and $\mathcal{L}^{\boldsymbol{\rho}}\tilde{\boldsymbol{\pi}} = \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}$ is also in $\Pi(\boldsymbol{\theta})$. Since $\boldsymbol{\theta}$ is chosen arbitrarily, we can then construct a single-valued function $\tilde{\boldsymbol{\pi}} : \mathbb{R}^d \to \Delta^N$ such that $\tilde{\boldsymbol{\pi}}(\boldsymbol{\theta}) \in \Pi(\boldsymbol{\theta})$ and $\|\tilde{\boldsymbol{\pi}}(\boldsymbol{\theta})\|_0 \leq \operatorname{affdim}(\mathcal{L}^{\boldsymbol{\rho}}) + 1$. Noting that $\max_{t\in\widehat{\mathcal{Y}}} \tilde{\pi}_t(\boldsymbol{\theta}) > 1/[\operatorname{affdim}(\mathcal{L}^{\boldsymbol{\rho}}) + 1]$, we can complete the rest of the proof similarly to Theorem 13. ■

Remark 21 *Interestingly, Ramaswamy and Agarwal [73] also used the affine dimension of the loss matrix to study the dimension of convex surrogates, suggesting that this concept plays a important role in loss function design and deserves more attention.*

B.7 Proof of Theorem 17

Proof. First of all, we prove that $\operatorname{relint}(\operatorname{conv}(\boldsymbol{\rho}(\mathcal{Y})))$ is in the image of $\nabla\Omega_T^*$ by contradiction. According to Rockafellar [79, Theorem 23.4], we have that $\partial\Omega_T(\boldsymbol{p})$ is non-empty for all $\boldsymbol{p} \in \operatorname{relint}(\operatorname{dom}(\Omega_T))$. Now suppose there exists $\boldsymbol{p} \in \operatorname{relint}(\operatorname{dom}(\Omega_T))$ such that $\boldsymbol{p} \neq \nabla\Omega_T^*(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \mathbb{R}^d$. Since $\partial\Omega_T(\boldsymbol{p})$ is non-empty, there exists $\boldsymbol{\theta'} \in \mathbb{R}^d$ such that $\boldsymbol{\theta'} = \nabla\Omega_T(\boldsymbol{p})$. By Rockafellar and Wets [80, Proposition 11.3] (applied on proper l.s.c. Ω_T), we have $\boldsymbol{p} = \nabla\Omega_T^*(\boldsymbol{\theta'})$, which contradicts the assumption $\boldsymbol{p} \neq \nabla\Omega_T^*(\boldsymbol{\theta'})$. Thus we have verified that $\operatorname{relint}(\operatorname{dom}(\Omega_T))$ is in the image of $\nabla\Omega_T^*$. This additionally implies that $\operatorname{relint}(\operatorname{conv}(\boldsymbol{\rho}(\mathcal{Y}))) \subseteq \operatorname{relint}(\operatorname{dom}(\Omega_T))$ is also in the image of $\nabla\Omega_T^*$ because of Condition 1.

Now we fix $\boldsymbol{\eta} \in \operatorname{relint}(\Delta^K)$ and note that

$$R_{L_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \Omega_T^*(\boldsymbol{\theta}) + \mathbb{E}_{y\sim\boldsymbol{\eta}}[\Omega_T(\boldsymbol{\rho}(y))] - \langle \boldsymbol{\theta}, \mathbb{E}_{y\sim\boldsymbol{\eta}}[\boldsymbol{\rho}(y)]\rangle$$

is differentiable and convex in $\boldsymbol{\theta}$. If $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is the minimizer of $R_{L_{\Omega_T}}(\cdot, \boldsymbol{\eta})$, its optimality condition indicates

$$\nabla_{\boldsymbol{\theta}} R_{L_{\Omega_T}}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) = \nabla\Omega_T^*(\boldsymbol{\theta}^*) - \mathbb{E}_{y\sim\boldsymbol{\eta}}[\boldsymbol{\rho}(y)] = 0.$$

When $\boldsymbol{\eta} \in \operatorname{relint}(\Delta^K)$, we have $\mathbb{E}_{y\sim\boldsymbol{\eta}}[\boldsymbol{\rho}(y)] \in \operatorname{relint}(\operatorname{conv}(\boldsymbol{\rho}(\mathcal{Y})))$, and thus $\boldsymbol{\theta}^*$ satisfying the above optimality condition does exist. Finally, we conclude the proof by combining Lemmas 8 and 12. ■

C Deferred Proofs in Section 4

C.1 Proof of Lemma 18

Proof. Denote by $V_{\boldsymbol{\theta}}(\boldsymbol{\pi}) \coloneqq \sum_{i=1}^K e^{\theta_i+1-\pi_i}$. Since $\ln(\cdot)$ is strictly increasing on $\mathbb{R}_{>0}$, $V_{\boldsymbol{\theta}}(\boldsymbol{\pi})$ shares the same minimizer as $\ln\left(\sum_{i=1}^K e^{\theta_i+1-\pi_i}\right)$. The Lagrangian $\mathcal{F}$ of the minimization problem $V_{\boldsymbol{\theta}}$ for $\boldsymbol{\pi} \in \Delta^K$ is written as follows:

$$\mathcal{F}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \beta) = \sum_{i=1}^K e^{\theta_i+1-\pi_i} - \boldsymbol{\alpha}^\top\boldsymbol{\pi} + \beta\left(\sum_{i=1}^K \pi_i - 1\right),$$

where $\alpha_1, \ldots, \alpha_K \geq 0$ and β are the Lagrangian multipliers. Since $V_{\boldsymbol{\theta}}(\boldsymbol{\pi})$ is convex and differentiable, and the feasible region Δ^K is convex, the KKT conditions are necessary and sufficient for the optimality of $\boldsymbol{\pi}^* \in \Pi(\boldsymbol{\theta})$, which requires that there exists $(\boldsymbol{\alpha}^*, \beta^*)$ satisfying

$$-e^{\theta_y+1-\pi_y^*} - \alpha_y^* + \beta^* = 0, \qquad \text{for any } y = 1, \ldots, K, \tag{20}$$

$$\mathbf{1}^\top\boldsymbol{\pi}^* = 1, \quad \boldsymbol{\pi}^* \geq 0, \quad \boldsymbol{\alpha}^* \geq 0, \tag{21}$$

$$\alpha_y^*\pi_y^* = 0, \qquad \text{for any } y = 1, \ldots, K. \tag{22}$$

The conditions (21) and (22) indicates that $\pi_y^* > 0 \implies \alpha_y^* = 0$. Then according to (20), we have

$$\pi_y^* > 0 \implies \pi_y^* = \theta_y - (\ln\beta^* - 1). \tag{23}$$

Meanwhile, for $\pi_y^* = 0$, (22) indicates that $\alpha_y^* \geq 0$, which further indicates

$$\pi_y^* = 0 \implies \theta_y = \ln(\beta^* - \alpha_y^*) - 1 \leq \ln \beta^* - 1. \tag{24}$$

Then (23) and (24) can be rewritten as follows:

$$\pi_y^* = \max\{\theta_y - (\ln \beta^* - 1), 0\}. \tag{25}$$

According to (21) and $\ln \beta^* - 1 \in \mathbb{R}$, the KKT conditions can be further simplified into the existence of $\tau^* \in \mathbb{R}$ such that the following conditions simultaneously hold:

$$\begin{cases} \pi_y^* = \max\{\theta_y - \tau^*, 0\}, \\ \sum_{y=1}^K \max\{\theta_y - \tau^*, 0\} = 1. \end{cases}$$

Denote by $f(\tau) \coloneqq \sum_{y=1}^K \max\{\theta_y - \tau, 0\}$. Then f is continuous on $\mathbb{R}$, and moreover, it is strictly decreasing on $(-\infty, \theta_{(1)}]$ and equals 0 for $\tau \geq \theta_{(1)}$ because of the following expression:

$$f(\tau) = \begin{cases} \sum_{y=1}^K (\theta_y - \tau) & \text{if } \tau \leq \theta_{(K)}, \\ \sum_{i=1}^{k-1} (\theta_{(i)} - \tau) & \text{if } \tau \in [\theta_{(k)}, \theta_{(k-1)}] \text{ for } k = 2, \ldots, K, \\ 0 & \text{if } \tau > \theta_{(1)}. \end{cases}$$

Then n defined in line 2 of Algorithm 1 exists since $f(\theta_{(1)}) = 0 < 1$. We also have that $f(\theta_{(n)}) = \sum_{i=1}^n (\theta_{(i)} - \theta_{(n)}) < 1$. When $n < K$, we have that $f(\theta_{(n+1)}) = \sum_{i=1}^{n+1} (\theta_{(i)} - \theta_{(n+1)}) = \sum_{i=1}^n (\theta_{(i)} - \theta_{(n+1)}) \geq 1$ by definition. Then we see

$$\frac{\sum_{i=1}^n \theta_{(i)} - 1}{n} \in [\theta_{(n+1)}, \theta_{(n)}]$$

and

$$\begin{aligned} f\left(\frac{\sum_{i=1}^n \theta_{(i)} - 1}{n}\right) &= \sum_{i=1}^n \left(\theta_{(i)} - \frac{\sum_{i=1}^n \theta_{(i)} - 1}{n}\right) \\ &= \sum_{i=1}^n \theta_{(i)} - \sum_{i=1}^n \theta_{(i)} + 1 \\ &= 1. \end{aligned}$$

When $n = K$, we have that $\sum_{i=1}^K \theta_{(i)} - K\theta_{(k)} < 1$, that is,

$$\frac{\sum_{i=1}^K \theta_{(i)} - 1}{K} < \theta_{[K]}.$$

Then

$$\begin{aligned} f\left(\frac{\sum_{i=1}^K \theta_{(i)} - 1}{K}\right) &= \sum_{i=1}^K \left(\theta_{(i)} - \frac{\sum_{i=1}^K \theta_{(i)} - 1}{K}\right) \\ &= \sum_{i=1}^K \theta_{(i)} - \sum_{i=1}^K \theta_{(i)} + 1 \\ &= 1. \end{aligned}$$

Combining these two cases, we have

$$\tau^* = \frac{\sum_{i=1}^n \theta_{(i)} - 1}{n} < \theta_{(1)},$$

which is what line 3 of Algorithm 1 returns. Since $f(\tau)$ is strictly decreasing on $(-\infty, \theta_{(1)}]$, τ^* uniquely satisfies $f(\tau^*) = 1$.

Since $\sum_{i=1}^k \theta_{(i)}$ can be calculated cumulatively in $\mathcal{O}(K)$ time and the quick sort runs in $\mathcal{O}(K \log K)$ time, Algorithm 1 runs in $\mathcal{O}(K \log K)$ time in total. Furthermore, since τ^* is unique and $\pi_y^* = \max\{\theta_y - \tau^*, 0\}$, we ensure that $\Pi(\boldsymbol{\theta})$ is a singleton, which concludes the proof. ■

C.2 Proof of Proposition 19

Proof. By Definition 10, the statement is equivalent to

$$\operatorname*{argmax}_{t\in\mathcal{Y}} \theta_t = \operatorname*{argmax}_{t\in\mathcal{Y}} \pi_{\log}(\boldsymbol{\theta})_t.$$

According to Lemma 18 and Algorithm 1, there exists $\tau \in \mathbb{R}$ such that $\pi_{\log}(\theta)_i = \max\{\theta_i - \tau, 0\}$. We also have that $\max_t \theta_t > \tau$ by contradiction: if $\max_t \theta_t \leq \tau$, $\boldsymbol{\pi}_{\log}(\boldsymbol{\theta}) = \mathbf{0}$ holds, which contradicts $\boldsymbol{\pi}_{\log}(\boldsymbol{\theta}) \in \Delta^K$.

Denote the set $\operatorname{argmax}_{t\in\mathcal{Y}}\theta_t$ by $\mathcal{I}$. For any $i \in \mathcal{I}$, we have $\pi_{\log}(\boldsymbol{\theta})_i = \max\{\theta_i - \tau, 0\} = \max_t \theta_t - \tau > 0$. For any $j \notin \mathcal{I}$,

- If $\theta_j > \tau$: $\pi_{\log}(\boldsymbol{\theta})_j = \max\{\theta_j - \tau, 0\} = \theta_j - \tau < \max_t \theta_t - \tau$,
- If $\theta_j \leq \tau$: $\pi_{\log}(\boldsymbol{\theta})_j = \max\{\theta_j - \tau, 0\} = 0 < \max_t \theta_t - \tau$.

These imply that for any $i \in \mathcal{I}$, $\pi_{\log}(\boldsymbol{\theta})_i \geq \pi_{\log}(\boldsymbol{\theta})_j$ and the equality holds if and only if $j \in \mathcal{I}$, which concludes the proof. ■

D Additional Examples

In this section, we further provide more examples of target losses to demonstrate that we can generate convex smooth surrogate losses for a wide range of target prediction problems. The generated convolutional Fenchel–Young losses are automatically guaranteed to entail linear surrogate regret bounds thanks to Theorem 13.

D.1 Multiclass Classification with Rejection

Problem setup. In multiclass classification with rejection [25], the class space is $\mathcal{Y} = [K]$, and the prediction space $\widehat{\mathcal{Y}} = [K+1]$, which is augmented by a rejection option $K+1$. We focus on the case that rejection cost c is in $[0, 0.5)$ here. For an input instance with class probability $\boldsymbol{\eta} \in \Delta^K$, the goal is to predict the most likely class label $t \in \operatorname{argmax}_{t\in\mathcal{Y}} \eta_t$ if $\max_{y\in\mathcal{Y}} \eta_y > 1 - c$ for the predetermined cost c, and refrain from predicting otherwise. The standard target loss is the 0-1-c loss:

$$\ell_{01c}(t, y) = \begin{cases} [\![t \neq y]\!], & t \in [K], \\ c, & t = K+1, \end{cases}$$

that is, the prediction suffers from the ordinary classification error if it is a wrong class label, and suffers from an intermediate error c if it chooses to refrain from prediction.

We adopt the decomposition (3) for this task: $\boldsymbol{\ell}^{\boldsymbol{\rho}}(t) = [\ell_{01c}(t, 1), \cdots, \ell_{01}(t, K)] = \mathbf{1} - \boldsymbol{e}_t$ if $t \neq K+1$, and $\boldsymbol{\ell}^{\boldsymbol{\rho}}(t) = c\mathbf{1}$ if $t = K+1$. We choose the label encoding function $\boldsymbol{\rho}(y)$... as in the multiclass classification task. In this case, $N = K+1$ and $d = K$, and $\mathcal{L}^{\boldsymbol{\rho}} = [\mathbf{1}\mathbf{1}^\top - I, c\mathbf{1}]$, where $\mathbf{1}$ is K-dimensional.

Loss formulation and calculation. In this case, we consider the Shannon negentropy Ω as in Section 4. Based on the discussion above and Lemma 8, the conjugated convolutional negentropy Ω_T^* can be written as follows:[6]

$$\Omega_T^*(\boldsymbol{\theta}) = \min_{\boldsymbol{\pi}\in\Delta^{K+1}} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}}\boldsymbol{\pi}) = \min_{\boldsymbol{\pi}\in\Delta^{K+1}} \ln\left(\sum_{i=1}^{K} \exp(\theta_i + 1 - \pi_i - (1-c)\pi_{K+1})\right), \quad (26)$$

and we can then get the corresponding convolutional Fenchel–Young loss as follows, by noting $\Omega_T(\boldsymbol{\rho}(y)) = 0$ for all $y \in \mathcal{Y}$:

$$L_{\Omega_T}(\boldsymbol{\theta}, y) = \min_{\boldsymbol{\pi}\in\Delta^{K+1}} \ln\left(\sum_{i=1}^{K} \exp(\theta_i + 1 - \pi_i - (1-c)\pi_{K+1})\right) - \theta_y. \quad (27)$$

To compute L_{Ω_T}, we need to know the minimizer $\boldsymbol{\pi}$ in (26). We show its closed form below.

Lemma 22 *For any $\boldsymbol{\theta} \in \mathbb{R}^K$, the problem* (26) *has a unique minimizer $\boldsymbol{\pi}^*(\boldsymbol{\theta}) \in \Pi(\boldsymbol{\theta})$, which can be obtained in $\mathcal{O}(K)$ time. Denote by y^* an arbitrary element in* $\operatorname{argmax}_{y'\in[K]} \theta_{y'}$,[7] *then the minimizer*

[6] We emphasize that the domain of conjugate is K-dimensional, despite the $\boldsymbol{\pi}$ is in $K+1$-simplex.

[7] Though chosen arbitrarily, uniqueness is guaranteed: argmax sets are singleton in the first and third cases.

can be written as follows:

$$\boldsymbol{\pi}^*(\boldsymbol{\theta}) = \begin{cases} \boldsymbol{e}_{y^*}, & \text{if } \frac{\exp(\theta_{y^*})}{\exp(\theta_{y^*}) + \sum_{i=1, i\neq y^*}^K \exp(\theta_i + 1)} > 1 - c \\ \boldsymbol{e}_{K+1}, & \text{if } \frac{\exp(\theta_{y^*})}{\sum_{i=1}^K \exp(\theta_i)} < 1 - c \\ \gamma(\boldsymbol{\theta})\boldsymbol{e}_{y^*} + (1 - \gamma(\boldsymbol{\theta}))\boldsymbol{e}_{K+1}, & \text{else,} \end{cases}$$

where $\gamma(\boldsymbol{\theta}) \coloneqq -\ln\left(\frac{\sum_{i=1}^K \exp(\theta_i)}{\exp(\theta_{y^*})} - 1\right) - \ln\left(\frac{1-c}{c}\right)$.

Proof. First, the Lagrangian of the minimization problem (26) is written as follows:

$$\begin{aligned} \mathcal{F}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \beta) &= \ln\left(\sum_{i=1}^K \exp(\theta_i + 1 - \pi_i - (1-c)\pi_{K+1})\right) - \boldsymbol{\alpha}^\top \boldsymbol{\pi} + \beta\left(\sum_{i=1}^{K+1} \pi_i - 1\right) \\ &= \ln\left(\sum_{i=1}^K \exp(\theta_i + 1 - \pi_i)\right) - (1-c)\pi_{K+1} - \boldsymbol{\alpha}^\top \boldsymbol{\pi} + \beta\left(\sum_{i=1}^{K+1} \pi_i - 1\right), \end{aligned}$$

where $\alpha_1, \ldots, \alpha_{K+1} \geq 0$ and $\beta \in \mathbb{R}$ are the Lagrangian multipliers. Since the log-sum-exp term is convex and differentiable and the feasible region Δ^{K+1} is convex, the KKT conditions are necessary and sufficient for the optimality of $\boldsymbol{\pi}^* \in \Pi(\boldsymbol{\theta})$, which requires that there exists $(\boldsymbol{\alpha}^*, \beta^*)$ satisfying

$$\frac{\exp(\theta_y + 1 - \pi_y^*)}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi_i^*)} = \beta^* - \alpha_y^*, \qquad \text{for any } y = 1, \ldots, K, \tag{28}$$

$$1 - c = \beta^* - \alpha_{K+1}^*, \tag{29}$$

$$\mathbf{1}^\top \boldsymbol{\pi}^* = 1, \quad \boldsymbol{\pi}^* \geq 0, \quad \boldsymbol{\alpha}^* \geq 0, \tag{30}$$

$$\alpha_y^* \pi_y^* = 0, \qquad \text{for any } y = 1, \ldots, K+1. \tag{31}$$

We continue the proof with the following four observations:

Observation 1. By noting $1 - c > 0.5$, we have

$$\frac{\exp(\theta_{y^*})}{\sum_{i=1}^K \exp(\theta_i)} \geq 1 - c \implies \operatorname*{argmax}_{y' \in [K]} \theta_{y'} = \{y^*\}.$$

Observation 2. From the KKT conditions, we have

$$\frac{\exp(\theta_{y^*})}{\exp(\theta_{y^*}) + \sum_{i=1, i\neq y^*}^K \exp(\theta_i + 1)} > 1 - c \implies \boldsymbol{\pi}^* = \boldsymbol{e}_{y^*}.$$

To see this, let us integrate the above left-hand side with (28) and (29):

$$\begin{aligned} \beta^* - \alpha_{y^*}^* &= \frac{\exp(\theta_{y^*} + 1 - \pi_{y^*}^*)}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi_i^*)} && \text{by (28)} \\ &= \frac{\exp(\theta_{y^*})}{\exp(\theta_{y^*}) + \sum_{i=1, i\neq y^*}^K \exp(\theta_i + \pi_{y^*}^* - \pi_i^*)} \\ &> \frac{\exp(\theta_{y^*})}{\exp(\theta_{y^*}) + \sum_{i=1, i\neq y^*}^K \exp(\theta_i + 1)} && \text{by } \pi_{y^*}^* \leq 1 \text{ and } \pi_i^* \geq 0 \\ &\overset{(\clubsuit)}{>} 1 - c && \text{by the assumption} \\ &= \beta^* - \alpha_{K+1}^*, && \text{by (29)} \end{aligned}$$

which indicates that $\alpha_{K+1}^* > \alpha_{y^*}^* \geq 0$, and thus $\pi_{K+1}^* = 0$ due to (30) and (31). On the other hand, for any $y \in [K] \setminus \{y^*\}$, we have

$$\beta^* - \alpha_y^* = \frac{\exp(\theta_y + 1 - \pi_y^*)}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi_i^*)} \qquad \text{by (28)}$$

$$
\begin{aligned}
&\leq 1 - \frac{\exp(\theta_{y^*} + 1 - \pi^*_{y^*})}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} \\
&\leq 1 - (1 - c) && \text{by the same argument as } (\clubsuit) \\
&= c \\
&< 1 - c && \text{by } c \in [0, 0.5) \\
&= \beta^* - \alpha^*_{K+1}, && \text{by (29)}
\end{aligned}
$$

and thus $\alpha^*_y > \alpha^*_{K+1} > 0$, which indicates that $\pi^*_y = 0$ except for $y = y^*$ due to (30) and (31). Thus, we verify $\boldsymbol{\pi}^* = \boldsymbol{e}_{y^*}$.

Observation 3. From the KKT conditions, we have

$$
\frac{\exp(\theta_{y^*})}{\sum_{i=1}^K \exp(\theta_i)} < 1 - c \implies \boldsymbol{\pi}^* = \boldsymbol{e}_{K+1}.
$$

To see this, we show the contraposition of Observation 3. Suppose there exists $y \in [K]$ such that $\pi^*_y > 0$. Then, we have $\alpha^*_y = 0$ due to (31), and consequently

$$
\begin{aligned}
\frac{\exp(\theta_y + 1 - \pi^*_y)}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} &= \beta^* && \text{by (28) and } \alpha^*_y = 0 \\
&\geq \beta^* - \alpha^*_{K+1} && \text{by (30)} \\
&= 1 - c. && \text{by (29)}
\end{aligned} \tag{32}
$$

Then, for any $y' \in [K] \setminus \{y\}$, this inequality implies

$$
\begin{aligned}
\beta^* - \alpha^*_{y'} &= \frac{\exp(\theta_{y'} + 1 - \pi^*_{y'})}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} && \text{by (28)} \\
&\leq 1 - \frac{\exp(\theta_y + 1 - \pi^*_y)}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} \\
&\leq 1 - (1 - c) && \text{by (32)} \\
&= c \\
&< 1 - c, && \text{by } c \in [0.0.5)
\end{aligned}
$$

which indicates that $\alpha^*_{y'} > 0$ and $\pi^*_{y'} = 0$ due to (30) and (31). Then, we have

$$
\begin{aligned}
\frac{\exp(\theta_y + 1 - \pi^*_y)}{\exp(\theta_y + 1 - \pi^*_y) + \sum_{i=1, i\neq y}^K \exp(\theta_i + 1)} &= \frac{\exp(\theta_y - \pi^*_y)}{\exp(\theta_y - \pi^*_y) + \sum_{i=1, i\neq y}^K \exp(\theta_i)} \\
&= \frac{\exp(\theta_y)}{\exp(\theta_y) + \sum_{i=1, i\neq y}^K \exp(\theta_i + \pi^*_y)} \\
&\geq 1 - c,
\end{aligned}
$$

where we used (32) at the last line. This inequality further implies

$$
\frac{\exp(\theta_{y^*})}{\sum_{i=1}^K \exp(\theta_i)} \geq \frac{\exp(\theta_y)}{\exp(\theta_y) + \sum_{i=1, i\neq y}^K \exp(\theta_i + \pi^*_y)} \geq 1 - c.
$$

Thus, the contraposition of Observation 3 is shown.

Observation 4. From the KKT conditions again, we have

$$
\begin{aligned}
&\frac{\exp(\theta_{y^*})}{\sum_{i=1}^K \exp(\theta_i)} \geq 1 - c \geq \frac{\exp(\theta_{y^*})}{\exp(\theta_{y^*}) + \sum_{i=1, i\neq y^*}^K \exp(\theta_i + 1)} \\
&\qquad \implies \boldsymbol{\pi}^* = \gamma(\boldsymbol{\theta}) \boldsymbol{e}_{y^*} + (1 - \gamma(\boldsymbol{\theta})) \boldsymbol{e}_{K+1},
\end{aligned}
$$

where $\gamma(\boldsymbol{\theta}) = -\ln\left(\frac{\sum_{i=1}^K \exp(\theta_y)}{\exp(\theta_{y^*})} - 1\right) - \ln\left(\frac{1-c}{c}\right)$. To see this, we carefully examine $\boldsymbol{\pi}^*$. First, if there exists different $y',\ y'' \in [K]$ with $\pi_{y'}, \pi_{y''} > 0$, we have

$$
\beta^* = \frac{\exp(\theta_{y'} + 1 - \pi^*_{y'})}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} \qquad \text{by (28) and } \alpha^*_{y'} = 0 \text{ since } \pi_{y'} > 0
$$

$$
\begin{aligned}
&= \frac{\exp(\theta_{y''} + 1 - \pi^*_{y''})}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} && \text{by (28) and } \alpha^*_{y''} = 0 \text{ since } \pi_{y'} > 0 \\
&\geq \beta^* - \alpha^*_{K+1} && \text{by } \alpha^*_{K+1} \geq 0 \\
&= 1 - c && \text{by (29)} \\
&> 0.5,
\end{aligned}
$$

which is impossible since

$$
\frac{\exp(\theta_{y'} + 1 - \pi^*_{y'})}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} + \frac{\exp(\theta_{y''} + 1 - \pi^*_{y''})}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} \leq 1.
$$

This contradiction indicates that there exists at most one $y \in [K]$ such that $\pi^*_y > 0$. Second, we show $y \in [K]$ with $\pi^*_y > 0$ must be $y = y^*$. To see this, suppose there exists $y \in [K]$ such that $\pi^*_y > 0$ but $y \neq y^*$, then $\pi^*_{y'} = 0$ for any $y' \in [K] \setminus \{y\}$ and we have

$$
\begin{aligned}
\frac{\exp(\theta_y + 1 - \pi^*_y)}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} &= \frac{\exp(\theta_y - \pi^*_y)}{\exp(\theta_y - \pi^*_y) + \sum_{i=1, i\neq y}^K \exp(\theta_i)} \\
&= \beta^* - \alpha^*_y && \text{by (28)} \\
&= \beta^* && \text{by } \pi^*_y > 0 \text{ and (31)} \\
&\geq \beta^* - \alpha^*_{K+1} && \text{by } \alpha^*_{K+1} \geq 0 \\
&\geq 1 - c && \text{by (29)} \\
&> 0.5.
\end{aligned}
$$

However, this contradicts the following inequality:

$$
\begin{aligned}
\frac{\exp(\theta_{y^*} + 1 - \pi^*_{y^*})}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} &= \frac{\exp(\theta_{y^*})}{\exp(\theta_y - \pi^*_y) + \sum_{i=1, i\neq y}^K \exp(\theta_i)} \\
&\geq \frac{\exp(\theta_{y^*})}{\sum_{i=1}^K \exp(\theta_i)} && \text{by } \pi^*_y \geq 0 \\
&\geq 1 - c && \text{by the assumption} \\
&> 0.5.
\end{aligned}
$$

Thus, we have verified that $\pi^*_y > 0$ is possible with $y = y^*$ or $y = K + 1$ only. From this, we have

$$
\boldsymbol{\pi}^* = \pi^*_{y^*} \boldsymbol{e}_{y^*} + (1 - \pi^*_{y^*}) \boldsymbol{e}_{K+1}. \tag{33}
$$

From now on, we show that

$$
\begin{cases}
\boldsymbol{\pi}^* = \gamma(\boldsymbol{\theta}) \boldsymbol{e}_{y^*} + (1 - \gamma(\boldsymbol{\theta})) \boldsymbol{e}_{K+1} \\
\boldsymbol{\alpha}^* = \mathbf{0} \\
\beta^* = 1 - c
\end{cases} \tag{34}
$$

fulfill the KKT conditions under the assumption of Observation 4. By using (33), we have

$$
\begin{aligned}
\frac{\exp(\theta_{y^*} + 1 - \pi^*_{y^*})}{\sum_{i=1}^K \exp(\theta_i + 1 - \pi^*_i)} &= \frac{\exp(\theta_{y^*} - \pi^*_{y^*})}{\exp(\theta_{y^*} - \pi^*_{y^*}) + \sum_{i=1, i\neq y^*}^K \exp(\theta_i)} && \text{by (33)} \\
&= \beta^* - \alpha^*_{y^*} && \text{by (28)} \\
&= 1 - c + \alpha^*_{y^*} - \alpha^*_{K+1}. && \text{by (29)}
\end{aligned}
$$

By elementary algebra and plugging in $\boldsymbol{\alpha}^* = \mathbf{0}$ given by (34), we can solve it with respect to $\pi^*_{y^*}$ as follows:

$$
\pi^*_{y^*} = -\ln \left(\frac{\sum_{i=1}^K \exp(\theta_i)}{\exp(\theta_{y^*})} - 1 \right) - \ln \left(\frac{1 - c}{c} \right) = \gamma(\boldsymbol{\theta}). \tag{35}
$$

By rearranging the assumption of Observation 4, we have

$$\frac{1}{e}\frac{c}{1-c} \leq \frac{\sum_{i=1}^{K}\exp(\theta_i)}{\exp(\theta_{y^*})} - 1 \leq \frac{c}{1-c},$$

from which we can see $0 \leq \pi^*_{y^*} = \gamma(\boldsymbol{\theta}) \leq 1$. All in all, $(\boldsymbol{\pi}^*, \boldsymbol{\alpha}^*, \beta^*)$ in (34) fulfill the KKT conditions, and thus Observation 4 is verified.

By combining Observations 2, 3, and 4, we have shown the closed form of $\pi^*(\boldsymbol{\theta})$. Furthermore, Observation 1 guarantees the uniqueness of $\boldsymbol{\pi}^*(\boldsymbol{\theta})$.

Lastly, it is easy to see that $\boldsymbol{\pi}^*(\boldsymbol{\theta})$ can be computed in linear time because both of the following terms

$$\frac{\exp(\theta_{y^*})}{\exp(\theta_{y^*}) + \sum_{i=1, i\neq y^*}^{K}\exp(\theta_i + 1)} \quad \text{and} \quad \frac{\exp(\theta_{y^*})}{\sum_{i=1}^{K}\exp(\theta_i)}$$

can be computed in linear time. ■

D.2 Multilabel Learning with Precision@k

Problem setup. In multilabel learning, the target prediction space $\mathcal{Y} = [2^d]$ is the collection of indices of all possible combinations of d binary labels with $|\mathcal{Y}| = K = 2^d$. Precision@k is a common performance metric for multilabel ranking. We consider that the prediction space $\widehat{\mathcal{Y}} = [\binom{d}{k}]$ is the collection of indices of all possible size-k subsets of multilabels with $|\widehat{\mathcal{Y}}| = N = \binom{d}{k}$. In addition, we encode the label $y \in \mathcal{Y}$ into $\boldsymbol{\rho}(y) \in \{0,1\}^d$ and the prediction $t \in \widehat{\mathcal{Y}}$ into $\boldsymbol{\mu}(t) \in \{0,1\}^d$, where $\{\boldsymbol{\mu}(t)\}_{t=1}^{N}$ is the collection of all distinct permutations of $\boldsymbol{\omega} = \boldsymbol{e}_1 + \cdots + \boldsymbol{e}_k$. Then, the target loss of Precision@k is defined as follows:

$$\ell(t, y) = 1 - \frac{\sum_{i=1}^{d}\rho(y)_i\mu(t)_i}{k}.$$

This is the portion of binary labels with value 0 in the top-k list.

We adopt the decomposition (3) by using the aforementioned $\boldsymbol{\rho}(y)$, $\boldsymbol{\ell}^{\rho}(t) = -\frac{\boldsymbol{\mu}(t)}{k}$, and $c(y) = 1$ for all $y \in \mathcal{Y}$.

Loss formulation and calculation. For Precision@k, we consider the base negentropy

$$\Omega(\boldsymbol{p}) = \sum_{i=1}^{d}(p_i \ln p_i + (1-p_i)\ln(1-p_i)).$$

Then, we have

$$\Omega^*(\boldsymbol{\theta}) = \sum_{i=1}^{d}\ln(1+\exp(\theta_i)),$$

and

$$\Omega_T^*(\boldsymbol{\theta}) = \min_{\boldsymbol{\pi}\in\Delta^{\binom{d}{k}}} \sum_{i=1}^{d}\ln\left(1+\exp\left(\theta_i - \frac{1}{k}\sum_{t=1}^{\binom{d}{k}}\pi_t\mu(t)_i\right)\right). \tag{36}$$

We can simplify it into the following problem:

$$\Omega_T^*(\boldsymbol{\theta}) = \min_{\boldsymbol{v}\in V} \sum_{i=1}^{d}\ln\left(1+\exp\left(\theta_i - \frac{1}{k}v_i\right)\right), \tag{37}$$

where $V \coloneqq \{\boldsymbol{v} \in \mathbb{R}^d : v_i \in [0,1], \|\boldsymbol{v}\|_1 = k\}$. Then, the convolutional Fenchel–Young loss can be written as follows:

$$L_{\Omega_T}(\boldsymbol{\theta}, y) = \min_{\boldsymbol{v}\in V} \sum_{i=1}^{d}\ln\left(1+\exp\left(\theta_i - \frac{1}{k}v_i\right)\right) - \langle\boldsymbol{\rho}(y), \boldsymbol{\theta}\rangle.$$

To calculate the gradient of $L_{\Omega_T}(\cdot, y)$, we only need the solution $\boldsymbol{v}^*$ in the optimization problem (37), which can be obtained efficiently.

Algorithm 2 Exact Solution of (37) in $\mathcal{O}(d \ln d)$

1: Set $f(\lambda) = \sum_{i=1}^{d} \max\{0, \min\{\theta_i - \lambda, 1\}\}$
2: Sort $\tilde{\boldsymbol{\theta}} = [\boldsymbol{\theta}; \boldsymbol{\theta} - \mathbf{1}] \in \mathbb{R}^{2d}$ such that $\tilde{\theta}_{(1)} \geq \cdots \geq \tilde{\theta}_{(2d)}$.
3: $n \leftarrow \max\{n \in [2d] : \ f(\tilde{\theta}_{(n)}) < k\}$.
4: $\lambda^* \leftarrow \tilde{\theta}_{(n)} + \frac{k - f(\tilde{\theta}_{(n)})}{f(\tilde{\theta}_{(n+1)}) - f(\tilde{\theta}_{(n)})}(\tilde{\theta}_{(n+1)} - \tilde{\theta}_{(n)})$.
5: $v_i^* \leftarrow \max\{0, \min\{\theta_i - \lambda^*, 1\}\}$.

Lemma 23 *For any $\boldsymbol{\theta} \in \mathbb{R}^d$, the minimization problem* (37) *has a unique minimizer $\boldsymbol{v}^*$, which can be obtained in $\mathcal{O}(d \log d)$ time with Algorithm 2.*

Proof. The uniqueness can be seen by noting that the objective function is strictly convex and its domain V is compact and convex.

Let us analyze the Lagrangian of the minimization problem (37), which can be written as follows:

$$\mathcal{F}(\boldsymbol{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = \sum_{i=1}^{d} \ln\left(1 + \exp\left(\theta_i - \frac{1}{k} v_i\right)\right) - \boldsymbol{\alpha}^\top \boldsymbol{v} + \boldsymbol{\beta}^\top (\boldsymbol{v} - \mathbf{1}) + \gamma \left(\sum_{i=1}^{d} v_i - k\right), \quad (38)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^d_{\geq 0}$, $\boldsymbol{\beta} \in \mathbb{R}^d_{\geq 0}$, and $\gamma \in \mathbb{R}$ are the Lagrangian multipliers. Then, we have the following KKT conditions:

$$\frac{\exp(\theta_i - v_i^*)}{1 + \exp(\theta_i - v_i^*)} = \gamma^* + \beta_i^* - \alpha_i^*, \qquad \text{for any } i = 1, \ldots, d \quad (39)$$

$$\mathbf{1}^\top \boldsymbol{v}^* = k, \quad \mathbf{0} \leq \boldsymbol{v}^* \leq \mathbf{1}, \quad \boldsymbol{\alpha}^* \geq \mathbf{0}, \quad \boldsymbol{\beta}^* \geq \mathbf{0}, \quad (40)$$

$$\alpha_i^* v_i^* = 0, \quad \beta_i^*(1 - v_i^*) = 0, \qquad \text{for any } i = 1, \ldots, d \quad (41)$$

where the inequalities in (40) are element-wise.

First, we show that the KKT conditions are fulfilled by $v_i^* = \max\{0, \min\{\theta_i - \lambda^*, 1\}\}$ for each $i \in [d]$, where $\lambda^* := \ln(\gamma^*/(1 - \gamma^*))$. Fixing $i \in [d]$, we divide into cases. If $v_i^* > \theta_i - \lambda^*$, we have

$$\frac{\exp(\theta_i - v_i^*)}{1 + \exp(\theta_i - v_i^*)} < \gamma^*,$$

which implies $\beta_i^* - \alpha_i^* < 0$ by (39). Since $\beta_i^* > 0$ by (40), we have $\alpha_i^* > 0$, which implies $v_i^* = 0$ due to (41). If $v_i^* < \theta_i - \lambda^*$, we can show $v_i^* = 1$ similarly. Thus, we have

- If $v_i^* > \theta_i - \lambda^*$, then $v_i^* = 0$;
- If $v_i^* < \theta_i - \lambda^*$, then $v_i^* = 1$.

Moreover, let us divide into cases on $\theta_i - \lambda^*$. If $\theta_i - \lambda^* < 0$, we must have $v_i^* > \theta_i - \lambda^*$ since $v_i \in [0, 1]$ (due to the feasibility (40)), which yields $v_i^* = 0$. If $\theta_i - \lambda^* > 1$, we have $v_i^* = 1$ similarly. If $0 \leq \theta_i - \lambda^* \leq 1$, we have $v_i^* = \theta_i - \lambda^*$ otherwise it ends up with contradiction—say, supposing $v_i^* > \theta_i - \lambda^*$, then we have $0 = v_i^* > \theta_i - \lambda^*$, which contradicts the premise $\theta_i - \lambda^* \in [0, 1]$. Combining all above, we have verified that

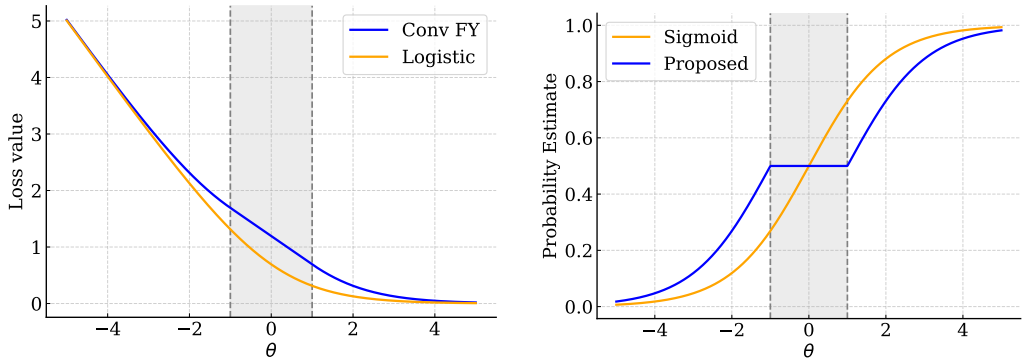
$$v_i^* = \max\{0, \min\{\theta_i - \lambda^*, 1\}\} \quad \text{for } i \in [d]$$

fulfills the KKT conditions.

Next, we show that Algorithm 2 returns this $\boldsymbol{v}^*$. By noting the constraint $\sum_{i=1}^{d} v_i^* = k$ in the feasibility conditions (40), we have

$$(f(\lambda^*) =) \sum_{i=1}^{d} \max\{0, \min\{\theta_i - \lambda^*, 1\}\} = k.$$

Thus, we need to solve $f(\lambda^*) = k$ to obtain $\boldsymbol{v}^*$. Let us sort the elements of $\tilde{\boldsymbol{\theta}} := [\boldsymbol{\theta}; \boldsymbol{\theta} - \mathbf{1}] \in \mathbb{R}^{2d}$ such that $\tilde{\theta}_{(1)} \geq \cdots \geq \tilde{\theta}_{[2d]}$. The solution to $f(\lambda^*) = k$ uniquely exists since we have



(a) Loss curves for $y = +1$.

(b) Probability estimators

Figure 1: Visualization of the convolutional Fenchel–Young loss, logistic loss, and probability estimators.

$f(\tilde{\theta}_{(1)}) = 0 < k$, $f(\tilde{\theta}_{[2d]}) = d > k$, and f is continuous and strictly decreasing on $[\tilde{\theta}_{[2d]}, \tilde{\theta}_{(1)}]$. From the strict decreasing nature, it can also be inferred that n in Step 3 of Algorithm 2 is the largest possible $n \in [2d]$ that $f(\tilde{\theta}_{(n)}) < k$ and $f(\tilde{\theta}_{(n+1)}) \geq k$, which indicates that the solution is in $(\tilde{\theta}_{n+1}, \tilde{\theta}_{(n)}]$. Furthermore, f is linear on each segment $[\tilde{\theta}_{[i+1]}, \tilde{\theta}_{(i)}]$. This indicates that for the point $\lambda^* \in [\tilde{\theta}_{(n+1)}, \tilde{\theta}_{(n)}]$, we have

$$f(\tilde{\theta}_{(n+1)}) + (\lambda^* - \tilde{\theta}_{(n)}) \frac{f(\tilde{\theta}_{(n+1)}) - f(\tilde{\theta}_{(n)})}{\tilde{\theta}_{(n+1)} - \tilde{\theta}_{(n)}} = f(\lambda^*) = k,$$

which yields Step 4 in Algorithm 2.

The time complexity of sorting $\tilde{\boldsymbol{\theta}}$ is $\mathcal{O}(d \log d)$. Note that the computation of $f(z)$ is in $\mathcal{O}(d)$ and n can be found using binary search due to monotonicity of f in $\mathcal{O}(\log d)$ steps, we can conclude that Step 3 is in $\mathcal{O}(d \log d)$. ■

E A Special Case: Visualization on Binary Classification

In this section, we visualize the graphs of convolutional Fenchel–Young losses and the corresponding probability estimator (provided in Theorem 17) for binary classification to get more intuition. For binary classification, we adopt the class space $\mathcal{Y} = \widehat{\mathcal{Y}} = \{-1, +1\}$, and the target loss $\ell_{01}(y, t) = [\![y \neq t]\!]$. We use the following decomposition of the target loss:

$$\rho(+1) = \frac{1}{2},\ \rho(-1) = -\frac{1}{2},\ \ell^{\rho}(+1) = -1,\ \ell^{\rho}(-1) = 1,\ \text{and}\ \ c(+1) = c(-1) = \frac{1}{2}.$$

With this decomposition, surrogate losses can operate on a univariate prediction, which is convenient for the visualization purpose.

We compare convolutional Fenchel–Young losses with the standard Fenchel–Young losses generated by the binary Shannon negentropy:

$$\Omega(p) = \left(\frac{1}{2} + p\right) \ln\left(\frac{1}{2} + p\right) + \left(\frac{1}{2} - p\right) \ln\left(\frac{1}{2} - p\right),$$

where $\frac{1}{2} + p =: \eta_{+1}$ is the positive class probability and $\frac{1}{2} - p =: \eta_{-1}$ is the negative class probability. The corresponding convolutional negentropy is as follows:

$$\Omega_T(p) = \left(\frac{1}{2} + p\right) \ln\left(\frac{1}{2} + p\right) + \left(\frac{1}{2} - p\right) \ln\left(\frac{1}{2} - p\right) + \max\{p, -p\},$$

where $\max\{p, -p\} = \frac{1}{2} - \min\{\eta_{+1}, \eta_{-1}\}$ is the negative Bayes 0-1 risk of class probability $(\eta_{+1}, \eta_{-1}) = (\frac{1}{2} + p, \frac{1}{2} - p)$. Then, the conjugate of convolutional negentropy can be written

as follows:

$$\Omega_T^*(\theta) = \begin{cases} \ln(1+\exp(\theta+1)) - \frac{\theta+1}{2}, & \text{if } \theta < -1 \\ \ln(2), & \text{if } -1 \le \theta \le 1 \\ \ln(1+\exp(\theta-1)) - \frac{\theta-1}{2}. & \text{if } 1 < \theta \end{cases}$$

Correspondingly, the convolutional Fenchel–Young loss is

$$L_{\Omega_T}(\theta, y) = \begin{cases} \ln(1+\exp(1+\theta)) - \frac{\theta(1+y)+1}{2}, & \text{if } \theta < -1 \\ \ln(2) - \frac{\theta y}{2}, & \text{if } -1 \le \theta \le 1 \\ \ln(1+\exp(\theta-1)) - \frac{\theta(1+y)-1}{2}. & \text{if } 1 < \theta \end{cases}$$

We can explicitly write down the gradient of the conjugated entropy as follows:

$$\nabla_\theta \Omega_T^*(\theta) = \begin{cases} \frac{\exp(1+\theta)}{1+\exp(1+\theta)} - \frac{1}{2}, & \theta < -1 \\ \frac{\exp(\theta-1)}{1+\exp(\theta-1)} - \frac{1}{2}, & \theta > 1 \\ 0, & \theta \in [-1, 1] \end{cases}$$

which is the estimator of $\eta_{+1}\rho(+1) + \eta_{-1}\rho(-1) = \frac{\eta_{+1}-\eta_{-1}}{2} = \frac{\eta_{+1}-1+\eta_{+1}}{2} = \eta_{+1} - \frac{1}{2}$ by Theorem 17. Finally, we can use the link function $\nabla_\theta \Omega_T^*(\theta) + \frac{1}{2}$ as the estimator of η_{+1}.

We show the convolutional Fenchel–Young loss and logistic loss (which is the standard Fenchel–Young loss generated by the binary Shannon negentropy) in Figure 1. It can be seen that the convolutional Fenchel–Young loss is linear in the shaded region $\theta \in [-1, 1]$, while resembling the logistic loss outside of this region. Compared with the sigmoid function used for probability estimation with logistic loss, the link function induced by the convolutional Fenchel–Young loss also generates a valid probability estimate in $[0, 1]$ for any $\theta \in \mathbb{R}$, with $\theta \in [-1, 1]$ generates constant value 0.5 as the estimate.

Further discussion. In case of binary classification, Figure 1 (a) nicely illustrates that the convolutional Fenchel–Young loss linearly "extends" the logistic loss at $\theta = 0$, which is the boundary point for binary classification. This illustration is possible because the above formulation for binary classification operates on the univariate score $\theta \in \mathbb{R}$. The linear extension at the boundary point aligns with Frongillo and Waggoner [39], which shows the square-root regret lower bound by assuming that a surrogate loss is locally strongly convex around the boundary points. In general prediction tasks, it is not straightforward to overcome the square-root regret lower bound by such a linear extension because the boundary points for a high-dimensional prediction task can be infinitely many. By contrast, convolutional Fenchel–Young losses provide a general recipe to get linear surrogate regret bound via infimal convolution.

F Additional Empirical Results

F.1 Multiclass Classification

To provide empirical validation of our proposed results, we use the loss introduced in Section 4 as an example and evaluate its performance on the ImageNet-1k dataset [32] using the ResNet-50 architecture [43].

Experimental Setup. We follow the default configuration of the PyTorch [71] ImageNet training script. Specifically, we use stochastic gradient descent (SGD) with a momentum of 0.9, training for 120 epochs with a mini-batch size of 256. The initial learning rate is set to 0.1 and divided by 10 every 30 epochs. For comparison, we also report results obtained using the standard cross-entropy loss under the same configuration. All experiments are conducted on 8 GeForce RTX 3090 GPUs, and we report the average validation accuracy over 3 independent runs.

Results and Discussion. As shown in Table 1, our proposed loss slightly outperforms the standard cross-entropy loss in terms of validation accuracy under a 5% t-test,. This improvement is achieved without any additional hyperparameter tuning, demonstrating the potential of our approach. Meanwhile, the computation time per epoch remains comparable between the two methods, which indicates the efficiency of the proposed loss. These results confirm the compatibility of our loss with both mini-batch and distributed optimization settings.

Table 1: Comparison between cross-entropy loss and the proposed loss on ImageNet-1k using ResNet-50.

Metric	**Cross-Entropy Loss**	**Proposed Loss** (18)
Accuracy (%)	76.40	**76.81**
Averaged Running Time / Epoch (s)	647.22	653.63

F.2 Classification with Rejection

We further evaluate the proposed loss (27) under the classification with rejection (CwR) setting, where the rejection cost is fixed at $c = 0.05$. The experiments are conducted on the CIFAR-10 and CIFAR-100 datasets [46]. We adopt the WideResNet-28 architecture [96] with a widen factor of 4 and a hidden dimension of 50 as the backbone network for all experiments.

Experimental Setup. We train each model for 120 epochs on 8 GeForce RTX 3090 GPUs with a per-GPU batch size of 128. The optimizer is SGD with a momentum of 0.9, weight decay of $5e-4$, and an initial learning rate of 0.1. The learning rate is scheduled by cosine annealing.

For data augmentation, each image is first converted to a tensor, then padded by 4 pixels on each side using reflection padding, followed by random cropping to 32×32 and random horizontal flipping. The images are finally normalized using the dataset-specific mean and standard deviation. We report the average system accuracy (1-averaged 0-1-c loss) and acceptance rate over 5 runs. Meanwhile, we also provide the result of ordinary classification using the similar loss (18) we proposed.

Table 2: CwR/Classification performance of proposed losses on CIFAR-10/100 using WideResNet-28.

Metric	**CIFAR-10**	**CIFAR-100**
CwR with (27)		
System Accuracy: c=0.05 (%)	96.79	91.90
Acceptance Rate (%)	92.84	65.94
Averaged Running Time / Epoch (s)	5.79	6.64
Classification with (18)		
Accuracy (%)	93.87	74.36
Averaged Running Time / Epoch (s)	6.07	7.41

Results and Discussion. As shown in Table 2, both CIFAR-10 and CIFAR-100 experiments demonstrate that our proposed loss successfully performs classification with rejection, achieving higher system accuracy while maintaining a reasonable acceptance rate. This indicates that the model learns to reject uncertain samples effectively without sacrificing overall performance. In addition, the average computation time per epoch is even lower than that of ordinary classification, which we attribute to the $\mathcal{O}(K)$ complexity of gradient computation for the proposed formulation, compared with the $\mathcal{O}(K \log K)$ cost in standard classification. These results confirm the efficiency and practicality of our loss for reliable decision-making under uncertainty.