Label Distributionally Robust Losses for Multi-class Classification: Consistency, Robustness and Adaptivity

Dixian Zhu [1] **Yiming Ying** [2] **Tianbao Yang** [3]

Abstract

We study a family of loss functions named label-distributionally robust (LDR) losses for multi-class classification that are formulated from distributionally robust optimization (DRO) perspective, where the uncertainty in the given label information are modeled and captured by taking the worse case of distributional weights. The benefits of this perspective are several fold: (i) it provides a unified framework to explain the classical cross-entropy (CE) loss and SVM loss and their variants, (ii) it includes a special family corresponding to the temperature-scaled CE loss, which is widely adopted but poorly understood; (iii) it allows us to achieve adaptivity to the uncertainty degree of label information at an instance level. Our contributions include: (1) we study both consistency and robustness by establishing top-k ($\forall k \geq 1$) consistency of LDR losses for multi-class classification, and a negative result that a top-1 consistent and symmetric robust loss cannot achieve top-k consistency simultaneously for all $k \geq 2$; (2) we propose a new adaptive LDR loss that automatically adapts the individualized temperature parameter to the noise degree of class label of each instance; (3) we demonstrate stable and competitive performance for the proposed adaptive LDR loss on 7 benchmark datasets under 6 noisy label and 1 clean settings against 13 loss functions, and on one real-world noisy dataset. The method is open-sourced at `https://github.com/Optimization-AI/ICML2023_LDR`.

[1]The University of Iowa, Iowa City, USA [2]University at Albany, Albany, USA [3]Texas A&M University, College Station, USA. Correspondence to: Dixian Zhu, Tianbao Yang <dixian-zhu@uiowa.edu, tianbao-yang@tamu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

In the past decades, multi-class classification has been used in a variety of areas, including image recognition (Krizhevsky et al., 2009; Deng et al., 2009; Geusebroek et al., 2005; Hsu & Lin, 2002), natural language processing (Lang, 1995; Brown et al., 2020; Howard & Ruder, 2018), etc. Loss functions have played an important role in multi-class classification (Janocha & Czarnecki, 2017; Kornblith et al., 2021).

In machine learning (ML), the most classical loss functions include the cross-entropy (CE) loss and the Crammer-Singer (CS) Loss (aka the SVM loss). Before the deep learning era, the CS loss have been widely studied and utilized for learning a linear or kernelized model (Crammer & Singer, 2002). However, in the deep learning era the CE loss seems to dominate the CS loss in practical usage (He et al., 2016). One reason is that the CE loss is consistent for multi-class classification (i.e., yielding Bayes optimal classifier with infinite number of samples), while the CS loss is generally not consistent unless under an exceptional condition that the maximum conditional probability of a class label given the input is larger than 0.5 (Zhang, 2004), which might hold for simple tasks such as MNIST classification but unlikely holds for complicated tasks, e.g., ImageNet classification (Beyer et al., 2020; Tsipras et al., 2020). Nevertheless, the CE loss has been found to be vulnerable to noise in the labels (Ghosh et al., 2017; Wang et al., 2019). Hence, efforts have been devoted to robustifying the CE loss such that it becomes noise-tolerant. A sufficient condition for the loss function to be noise-tolerant is known as the symmetric property (Ghosh et al., 2017), i.e., the sum of losses over all possible class labels for a given data point is a constant. Following the symmetric property, multiple variants of CE loss have been proposed to enjoy the noise-tolerance property. However, a symmetric loss alone is usually not enough to achieve good performance in practice due to slow convergence (Zhang & Sabuncu, 2018; Ma et al., 2020). Hence, a loss that interpolates between the CE loss and a symmetric loss usually achieves better performance on real-world noisy datasets (Wang et al., 2019; Zhang & Sabuncu, 2018; Englesson & Azizpour, 2021; Ma et al., 2020), e.g., on the WebVision data with 20% noisy data (Li

et al., 2017).

These existing results can be explained from the degree of uncertainty of the given label information. If the given label information of an instance is certain (e.g., MNIST data), then the CS loss is a good choice (Tang, 2013); if the given label information is little uncertain (e.g., Imagenet Data), then the CE loss is a better choice; if the given label information is highly uncertain (e.g., WebVision data), then a loss that interpolates between the CE loss and a symmetric loss wins. While the parameter tuning (e.g., the weights for combining the CE loss and a symmetric loss) can achieve dataset adaptivity, an open problem exists: "*Can we define a loss to achieve instance adaptivity, i.e., adaptivity to the uncertainty degree of label information of each instance?*"

To address this question, this paper proposes to capture the uncertainty in the given label information of each instance by using the distributionally robust optimization (DRO) framework. For any instance, we assign each class label a distributional weight and define the loss in the worst case of the distributional weights properly regularized by a function, which is referred to as label distributionally robust (LDR) losses. The benefits of using the DRO framework to model the uncertainty of the given label information include: (i) it provides a unified framework of many existing loss functions, including the CE loss, the CS loss, and some of their variants, e.g., top-k SVM loss. (ii) it includes an interesting family of losses by choosing the regularization function of the distributional weights as Kullback–Leibler (KL) divergence, which is referred to as LDR-KL loss for multi-class classification. The LDR-KL loss is closely related to the temperature-scaled CE loss that has been widely used in other scenarios (Hinton et al., 2015; Chen et al., 2020), where the temperature corresponds to the regularization parameter of the distribibutional weights. Hence, our theoretical analysis of LDR-KL loss can potentially shed light on using the temperature-scaled CE loss in different scenarios. (iii) The LDR-KL loss can interpolate between the CS loss (at one end), the (margin-aware) CE loss (in between) and a symmetric loss (at the other end) by varying the regularization parameter, hence providing more flexibility for adapting to the noisy degree of given label information. Our contributions are the following:

- We establish a strong consistency of the LDR-KL loss namely All-k consistency that is top-k consistency (corresponding to the Bayes optimal top-k error) for all $k \geq 1$, and a sufficient condition for the general family of LDR losses to be All-k consistent.
- We show that an extreme case of the LDR-KL loss satisfies the symmetric property, making it tolerant to label noise. In addition, we establish a negative result that a non-negative symmetric loss function cannot achieve top-k consistency for $k = 1$ and $k = 2$ simultaneously. Hence, we conclude that existing top-1 consistent symmetric losses cannot be All-k consistent.
- We propose an adaptive LDR-KL loss, which allows us to automatically learn the personalized temperature parameter of each instance adapting to its own noisy degree. We develop an efficient algorithm for empirical risk minimization based on the adaptive LDR-KL loss without incurring much additional computational burden.
- We conduct extensive experiments on multiple datasets under different noisy label settings, and demonstrate the stable and competitive performance of the proposed adaptive LDR-KL loss in all settings.

2. Related Work

Traditional multi-class loss functions. Multi-class loss functions have been studied extensively in conventional ML literature (Weston & Watkins, 1998; Tewari & Bartlett, 2007; Zhang, 2004; Crammer & Singer, 2002). Besides the CE loss, well-known loss functions include the CS loss and the Weston-Watkins (WW) loss. Consistency has been studied for these loss for achieving Bayes optimal solution under infinite amount of data. The CE loss is shown to be consistent (producing Bayes optimal classifier for minimizing zero-one loss, i.e., top-1 error), while the CS and WW loss are shown to be inconsistent (Zhang, 2004). Variants of these standard losses have been developed for minimizing top-k error ($k > 1$), such as (smoothed) top-k SVM loss, top-k CE loss (Lapin et al., 2015; 2016; 2018; Yang & Koyejo, 2020). Top-k consistency has been studied in these papers showing some losses are top-k consistent while other are not. However, the relationship between these losses and the CE loss is still unclear, which hinders the understanding of their insights. Hence, it is important to provide a unified view of these different losses, which could not only help us understand the existing losses but also provide new and potentially better choices.

Robust Loss functions. While may different methods have been developed for tackling label noise (Goldberger & Ben-Reuven, 2017; Han et al., 2018; Hendrycks et al., 2018; Patrini et al., 2017; Zhang et al., 2021; Madry et al., 2017; Zhang & Sabuncu, 2018; Wang et al., 2019), of most relevance to the present work are the symmetric losses and their combinations with the CE loss or its variants. Ghosh et al. (2017) established that a symmetric loss is robust against symmetric or uniform label noise where a true label is uniformly flipped to another class, i.e., the optimal model that minimizes the risk under the noisy data distribution also minimizes the risk under the true data distribution. When the optimal risk under the true data distribution is zero, the symmetric loss is also noise tolerant to the non-uniform label noise. The most well-known symmetric loss is the mean-

Table 1. Comparison between different loss functions. Notice that for SCE, we simplify the loss combination by convex combination ($\beta = 1 - \alpha$ and $0 \leq \alpha \leq 1$). Detailed expressions of different loss functions are included in the Appendix A. 'NA' means not available or unknown. * means that the results are derived by us (proofs are included in the Appendix J). Expressions in parentheses mean the conditions about the hyper-parameters under which the properties hold.

Category	Loss	Top-1 consistent	All-k consistent	Symmetric	instance adaptivity
Traditional	CE	Yes	Yes	No	No
	CS (Crammer & Singer, 2002)	No	No	No	No
	WW (Weston & Watkins, 1998)	No	No	No	No
Symmetric	MAE (RCE, clipped) (Ghosh et al., 2017)	Yes	No	Yes	No
	NCE (Ma et al., 2020)	Yes	No	Yes	No
	RLL (Patel & Sastry, 2021)	No	No	Yes	No
Interpolation between CE and MAE	GCE (Zhang & Sabuncu, 2018)	Yes	Yes* ($0 \leq q < 1$)	Yes ($q = 1$)	No
	SCE (Wang et al., 2019)	Yes	Yes* ($0 \leq \alpha < 1$)	Yes ($\alpha = 0$)	No
	JS (Englesson & Azizpour, 2021)	NA	NA	Yes ($\pi_1 = 1$)	No
Bounded	MSE (Ghosh et al., 2017)	Yes	Yes*	No	No
	TGCE (Zhang & Sabuncu, 2018)	NA	NA	No	No
Asymmetric	AGCE (Zhou et al., 2021)	Yes	No	No	No
	AUL (Zhou et al., 2021)	Yes	No	No	No
LDR	LDR-KL (this work)	Yes ($\lambda > 0$)		Yes ($\lambda = \infty$)	No
	ALDR-KL (this work)	Yes ($\lambda_0 > 0, \alpha > \frac{\log K}{\lambda_0}$)		Yes ($\lambda_0 = \infty$)	Yes

absolute-error (MAE) loss (aka unhinged loss) (van Rooyen et al., 2015). An equivalent (up to a constant) symmetric loss named the reverse-CE (RCE) loss is proposed (Wang et al., 2019), which reverses the predicted class probabilities and provided one-hot vector encoding the given label information in computing the KL divergence. Ma et al. (2020) show that a simple normalization trick can make any loss functions to be symmetric, e.g., normalized CE (NCE) loss. However, a symmetric loss alone might suffer from slow convergence issue for training (Zhang & Sabuncu, 2018; Wang et al., 2019; Ma et al., 2020). To address this issue, a loss that interpolates between the CE loss and a symmetric loss is usually employed. This includes the symmetric-CE (SCE) loss (Wang et al., 2019) that interpolates between the RCE loss and the CE loss, the generalized CE loss (GCE) that interpolates the MAE loss and the CE loss, Jensen-Shannon divergence (JS) loss that interpolates the MAE loss and the CE loss (Englesson & Azizpour, 2021). These losses differ in the way of interpolation. Ma et al. (2020) has generalized this principle to combine an active loss and a passive loss and proposed several variants of robust losses, e.g., NCE + MAE. Moreover, the symmetric property has been relaxed to bounded condition enjoyed by the robust logistic loss (RLL) (Patel & Sastry, 2021) and mean-square-error (MSE) loss (Ghosh et al., 2017), and also relaxed to asymmetric property (Zhou et al., 2021) enjoyed by a class of asymmetric functions. The asymmetric loss functions have been proved to be top-1 calibrated and hence top-1 consistent (Zhou et al., 2021). However, according to their definition the asymmetric losses cannot be top-k consistent for any $k \geq 2$ because it contradicts to the top-k calibration (Yang & Koyejo, 2020) - a necessary condition for top-k consistency. We provide a summary for the related robust and traditional multi-class loss functions in Table 4.

Distributional robust optimization (DRO) Objective. It is worth mentioning the difference from existing literature that leverages the DRO framework to model the distributional shift between the training data distribution and testing data distribution (Shalev-Shwartz & Wexler, 2016; Namkoong & Duchi, 2017; Qi et al., 2020; Levy et al., 2020; Duchi et al., 2016; Duchi & Namkoong, 2021). These existing works assign a distributional weight to each instance and define an aggregate objective over all instances by taking the worse case over the distributional weights in an uncertainty set that is formed by a constraint function explicitly or a regularization implicitly. In contrast, we study the loss function of an individual instance by using the DRO framework to model the uncertainty in the label information, which models a distributional shift between the true class distribution and the observed class distribution given an instance.

3. The LDR Losses

Notations. Let $(\mathbf{x}, y) \sim \mathbb{P}$ be a random data following underlying true distribution $\mathbb{P}$, where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes an input data, $y \in [K] := \{1, \ldots, K\}$ denotes its class label. Let $f(\mathbf{x}) \in \mathcal{C} \subset \mathbb{R}^K$ denote the prediction scores by a predictive function f on data $\mathbf{x}$. Let $f_k(\mathbf{x})$ denote the k-th entry of $f(\mathbf{x})$. Without causing any confusion, we simply use the notation $\mathbf{f} = (f_1, \ldots, f_K) = f(\mathbf{x}) \in \mathcal{C}$ to denote the prediction for any data $\mathbf{x}$. Interchangeably, we also use one-hot encoding vector $\mathbf{y} \in \mathbb{R}^K$ to denote the label information of a data. For any vector $\mathbf{q} \in \mathbb{R}^K$, let $q_{[k]}$ denote the k-largest entry of $\mathbf{q}$ with ties breaking arbitrarily. Let $\Delta_K = \{\mathbf{p} \in \mathbb{R}^K : \sum_{k=1}^K p_k = 1, p_k \geq 0\}$ denote a K-dimensional simplex. Let $\delta_{k,y}(f(\mathbf{x})) = f_k(\mathbf{x}) - f_y(\mathbf{x})$ denote the difference between two coordinates of the prediction scores $f(\mathbf{x})$.

3.1. Label-Distributionally Robust (LDR) Losses

A straightforward idea for learning a good prediction function is to enforce that f_y to be larger than any other coordinates of $\mathbf{f}$ with possibly a large margin. However, this approach could mis-guide the learning process due to that the label information $\mathbf{y}$ is often noisy, meaning that the zero entries in $\mathbf{y}$ are not necessarily true zero, the entry equal to one in $\mathbf{y}$ is not necessarily true one. To handle such uncertainty in the label information, we propose to use a regularized DRO framework to capture the inherent uncertainty in the label information. Specifically, the proposed family of LDR losses is defined by:

$$\psi_\lambda(\mathbf{x}, y) = \max_{\mathbf{p}\in\Omega} \sum_{k=1}^{K} p_k(\delta_{k,y}(f(\mathbf{x})) + c_{k,y}) - \lambda R(\mathbf{p}), \quad (1)$$

where $\mathbf{p}$ is referred to as Distributional Weight (DW) vector, $\Omega \subseteq \{\mathbf{p} \in \mathbb{R}^K : p_k \geq 0, \sum_k p_k \leq 1\}$ is a convex constraint set for the DW vector $\mathbf{p}$, $c_{k,y} = c_y \mathbb{I}(y \neq k)$ denotes the margin parameter with $c_y \geq 0$, $\lambda > 0$ is the DW regularization parameter, and $R(\mathbf{p})$ is a regularization term of $\mathbf{p}$ which is set by a **strongly convex function**. The strong convexity of $R(\mathbf{p})$ makes the LDR losses smooth in terms of $f(\mathbf{x})$ due to the duality between strongly convexity and smoothness (Kakade & Shalev-Shwartz, 2009; Nesterov, 2005). With a LDR loss, the risk minimization problem is formulated as following, where E denotes the expectation:

$$\min_{f:\mathcal{X}\to\mathbb{R}^K} L_\psi(f) := \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{P}}[\psi_\lambda(\mathbf{x}, y)]. \quad (2)$$

The LDR-KL Loss. A special loss in the family of LDR losses is called the LDR-KL loss defined by specifying $\Omega = \Delta_K$, and $R(\mathbf{p})$ as the KL divergence between $\mathbf{p}$ and the uniform probabilities $(1/K, \ldots, 1/K)$, i.e., $R(\mathbf{p}) = \mathrm{KL}(\mathbf{p}, 1/K) = \sum_{k=1}^K p_k \log(p_k K)$. In this case, we can derive a closed-form expression of $\psi_\lambda(\mathbf{x}, y)$, denoted by $\psi_\lambda^{\mathrm{KL}}(\mathbf{x}, y)$ (cf. Appendix B),

$$\psi_\lambda^{\mathrm{KL}}(\mathbf{x}, y) = \lambda \log\left[\frac{1}{K}\sum_{k=1}^{K} \exp\left(\frac{f_k(\mathbf{x}) + c_{k,y} - f_y(\mathbf{x})}{\lambda}\right)\right]. \quad (3)$$

It is notable that the DW regularization parameter λ is a key feature of this loss. In many literature, this parameter is called the temperature parameter, e.g., (Hinton et al., 2015; Chen et al., 2020). It is interesting to see that the LDR-KL loss covers several existing losses as special cases or extreme cases by varying the value of λ. We include the proof for the analytical forms of these cases in the appendix for completeness.

Extreme Case I. $\lambda = 0$. The loss function becomes: $\psi_0^{\mathrm{KL}}(\mathbf{x}, y) = \max(0, \max_{k\neq y} f_k(\mathbf{x}) + c_y - f_y(\mathbf{x}))$ which is known as Crammer-Singer loss (an extension of hinge loss) for multi-class SVM (Crammer & Singer, 2002).

Extreme Case II. $\lambda = \infty$. The loss function becomes: $\psi_\infty^{\mathrm{KL}}(\mathbf{x}, y) = \frac{1}{K}\sum_{k=1}^K (f_k(\mathbf{x}) - f_y(\mathbf{x}) + c_{k,y})$, which is similar to the Weston-Watkins (WW) loss (Weston & Watkins, 1998) given by $\sum_{k\neq y} \max(0, f_k(\mathbf{x}) - f_y(\mathbf{x}) + 1)$. However, WW loss is not top-1 consistent but the LDR-KL loss when $\lambda = \infty$ is top-1 consistent (Zhang, 2004).

A Special Case. When $\lambda = 1$, the LDR-KL loss becomes the CE loss (with a margin parameter): $\psi_1^{\mathrm{KL}}(\mathbf{x}, y) = \log\left(\sum_k \exp\left(f_k(\mathbf{x}) + c_{k,y} - f_y(\mathbf{x})\right)\right)$. It is notable that in standard CE loss, the margin parameter c_y is usually set to be zero. Adding a margin parameter makes the above CE loss enjoys the large-margin benefit. This also recovers the label-distribution-aware margin loss proposed by (Cao et al., 2019) for imbalanced data classification with different margin values c_y for different classes.

3.2. Consistency

Consistency is an important property of a loss function. In this section, we establish the consistency of LDR losses. A standard consistency is for achieving Bayes optimal top-1 error. We will show much stronger consistency for achieving Bayes optimal top-k error for any $k \in [K]$. To this end, we first introduce some definitions and an existing result of sufficient condition of top-k consistency by top-k calibration (Yang & Koyejo, 2020).

For any vector $\mathbf{f} \in \mathbb{R}^K$, we let $r_k(\mathbf{f})$ denote a top-k selector that selects the k indices of the largest entries of $\mathbf{f}$ by breaking ties arbitrarily. Given a data $(\mathbf{x}, y)$, its top-k error is defined as $\mathrm{err}_k(f, \mathbf{x}, y) = \mathbb{I}(y \notin r_k(f(\mathbf{x})))$. The goal of a classification algorithm under the top-k error metric is to learn a predictor f that minimizes the error: $L_{\mathrm{err}_k}(f) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{P}}[\mathrm{err}_k(f, \mathbf{x}, y)]$. The predictor $f^*(\cdot) : \mathcal{X} \to \mathbb{R}^K$ is top-k Bayes optimal if $L_{\mathrm{err}_k}(f^*) = L^*_{\mathrm{err}_k} := \inf_{f\in\mathcal{F}} L_{\mathrm{err}_k}(f)$. Since we can not directly optimize the top-k error, we usually optimize a risk function using a surrogate loss, e.g., (2). A central question of consistency is that whether optimizing the risk function can optimize the top-k error.

Definition 1. *For a fixed $k \in [K]$, a loss function ψ is top-k consistent if for any sequence of measurable functions $f^{(n)} : \mathcal{X} \to \mathbb{R}^K$, we have*

$$L_\psi(f^{(n)}) \to L^*_\psi \Longrightarrow L_{err_k}(f^{(n)}) \to L^*_{err_k}$$

where $L^*_\psi = \min_f L_\psi(f)$. If the above holds for all $k \in [K]$, it is referred to as All-k consistency.

According to (Yang & Koyejo, 2020), a necessary condition for top-k consistency of a multi-class loss function is that the loss function is top-k calibrated defined below. Since top-k calibration characterizes when minimizing ψ for a fixed $\mathbf{x}$ leads to the Bayes decision for that $\mathbf{x}$, we will consider $\mathbf{f} = f(\mathbf{x})$ for any fixed $\mathbf{x} \in \mathcal{X}$ and let $\mathbf{q} = \mathbf{q}(\mathbf{x}) = (\Pr(y = 1|\mathbf{x}), \ldots, \Pr(y = K|\mathbf{x}))$ be the underlying conditional label distribution. Define $L_\psi(\mathbf{f}, \mathbf{q}) = \sum_l q_l \psi(\mathbf{x}, l)$. Let $P_k(\mathbf{f}, \mathbf{q})$ denote that $\mathbf{f}$ is top-k preserving with respect to the underlying label distribution

$\mathbf{q}$, i.e., if for all $l \in [K]$, $q_l > q_{[k+1]} \implies f_l > f_{[k+1]}$, and $q_l < q_{[k]} \implies f_l < f_{[k]}$.

Definition 2. *For a fixed $k \in [K]$, a loss function ψ is called top-k calibrated if for all $\mathbf{q} \in \Delta_K$ it holds that: $\inf_{\mathbf{f}\in\mathbb{R}^K:\neg P_k(\mathbf{f},\mathbf{q})} L_\psi(\mathbf{f},\mathbf{q}) > \inf_{\mathbf{f}\in\mathbb{R}^K} L_\psi(\mathbf{f},\mathbf{q})$. A loss function is called All-k calibrated if the loss function ψ is top-k calibrated for all $k \in [K]$.*

Unlike (Yang & Koyejo, 2020) that relies on top-k preserving property of optimal predictions, we will use rank preserving property of optimal predictions to prove All-k calibration and consistency. $\mathbf{f}$ is called **rank preserving** w.r.t $\mathbf{q}$, i.e., if for any pair $q_i < q_j$ it holds that $f_i < f_j$. Our lemma below shows that if the optimal predictions for minimizing L_ψ is rank preserving, then ψ is All-k calibrated. All missing proofs are included in appendix.

Lemma 1. *For any $\mathbf{q} \in \Delta_K$, if $\mathbf{f}^* = \arg\min_{\mathbf{f}\in\mathbb{R}^K} L_\psi(\mathbf{f},\mathbf{q})$ is rank preserving with respect to $\mathbf{q}$, then ψ is All-k calibrated.*

Our first main result of consistency is the following.

Theorem 1. *If $c_y = c$, $\forall y \in [K]$, then $\mathbf{f}^* = \arg\min_{\mathbf{f}\in\mathbb{R}^K} L_{\psi_\lambda^{KL}}(\mathbf{f},\mathbf{q})$ is rank preserving for any $\lambda \in (0,\infty)$ and hence the LDR-KL loss ψ_λ^{KL} with any $\lambda \in (0,\infty)$ is All-k calibrated. Therefore, it is All-k consistent.*

Calibration of the general LDR losses: we present a sufficient condition on the DW regularization function R such that the resulting LDR loss ψ_λ is All-k calibrated. The result below will address an open problem in (Lapin et al., 2018) regarding the consistency of smoothed top-k SVM loss.

Definition 3. *A function $R(\mathbf{p})$ is input-symmetric if its value is the same, no matter the order of its arguments. A set Ω is symmetric, if a point $\mathbf{p}$ is in the set then so is any point obtained by interchanging any two coordinates of $\mathbf{p}$.*

Theorem 2. *If $c_y = c, \forall y \in [K]$, R and Ω are (input-) symmetric, R is strongly convex satisfying $\partial R(0) < 0$ and $[\partial R(p)]_i > 0 \implies p_i \neq 0$, then the family of LDR losses ψ_λ with any $\lambda \in (0,\infty)$ is All-k calibrated.*

Remark: Examples of R that satisfy the above conditions include $R(\mathbf{p}) = \sum_i p_i \log(Kp_i)$ and $R(\mathbf{p}) = \|\mathbf{p} - 1/K\|^2$.

Consistency of smoothed top-k SVM losses. As an application of our result above, we restrict our attention to $\Omega = \Omega(k) = \{\mathbf{p} \in \mathbb{R}^K : \sum_{l=1}^K p_l \leq 1, p_l \leq 1/k, \forall l\}$. When $\lambda = 0$, the LDR loss becomes the top-k SVM loss (Lapin et al., 2018), i.e.,

$$\psi_{\lambda=0}^k(\mathbf{x},y) = \max_{\mathbf{p}\in\Omega(k)} \sum_l p_l(\delta_{l,y}(f(\mathbf{x})) + c_{l,y})$$
$$= \frac{1}{k}\sum_{i=1}^k \max(0, \mathbf{f} - f_y + \mathbf{c}_y)_{[i]},$$

where $\mathbf{c}_y = (c_{1,y},\ldots,c_{K,y})$. This top-$k$ SVM loss is not consistent acoording to (Lapin et al., 2018). But, we can make it consistent by adding a strongly convex regularizer $R(\mathbf{p})$ satisfying the conditions in Theorem 2. This addresses the open problem raised in (Lapin et al., 2018) regarding the consistency of smoothed top-k SVM loss for multi-class classification.

3.3. Robustness

Definition 4. *A loss function ψ is symmetric if: $\sum_{j=1}^K \psi(\mathbf{x},j)$ is a constant for any $f,\mathbf{x}$.*

A loss with the symmetric property is noise-tolerant against uniform noise and class dependent noise under certain conditions (Ghosh et al., 2017). For example, under uniform noise if the probability of noise label $\hat{y}$ not equal to true label y is not significantly large, i.e., $\Pr(\hat{y} \neq y|y) < 1 - \frac{1}{K}$, then the optimal predictor for optimizing $\mathrm{E}_{\mathbf{x},\hat{y}}[\psi(\mathbf{x},\hat{y})]$ is the same as optimizing the true risk $\mathrm{E}_{\mathbf{x},y}[\psi(\mathbf{x},y)]$.

Next, we present a negative result showing that a non-negative symmetric loss cannot be All-k consistent.

Theorem 3. *A non-negative symmetric loss function cannot be top-$1,2$ consistent simultaneously when $K \geq 3$.*

Remark: The above theorem indicates that it is impossible to design a non-negative loss that is not only symmetric but also All-k consistent. Exemplar non-negative symmetric losses include MAE (or SCE or unhinged loss) and NCE, which are top-1 consistent but not All-k consistent.

Next, we show a way to break the impossibility by using a symmetric loss that is not necessarily non-negative, e.g., the LDR-KL loss with $\lambda = +\infty$.

Theorem 4. *When $\lambda = \infty$, the LDR-KL loss enjoys symmetric property, and is All-k consistent when $\mathcal{C} = \{\mathbf{f} \in \mathbb{R}^K, \|\mathbf{f}\|_2 \leq B\}$.*

Remark: Theorem 4 does not contradict to Theorem 3 because LDR-KL loss with $\lambda = \infty$ is not non-negative.

3.4. Adaptive LDR-KL (ALDR-KL) Loss

Although LDR-KL with $\lambda = \infty$ is both robust and consistent, it may suffer from the slow convergence issue similar to the MAE loss (Zhang & Sabuncu, 2018). We design an Adaptive LDR-KL (ALDR-KL) loss for the data with label noises, which can automatically adjust DW regularization parameter λ for individual data. The motivation is: i) for those noisy data, the model is less confident and we want to make the λ to be larger, so that loss function is more robust; ii) for those clean data, the model is more confident and we want to make the λ to be smaller to enjoy the large margin property of the CS loss.

To this end, we propose the following ALDR-KL loss:

$$\psi_{\alpha,\lambda_0}^{\mathrm{KL}}(\mathbf{x},y) = \max_{\lambda\in\mathbb{R}_+} \max_{\mathbf{p}\in\Delta_K} \sum_{k=1}^K p_k(\delta_{k,y}(f(\mathbf{x})) + c_{k,y})$$
$$- \lambda \mathrm{KL}(\mathbf{p}, \frac{1}{K}) - \frac{\alpha}{2}(\lambda - \lambda_0)^2. \quad (4)$$

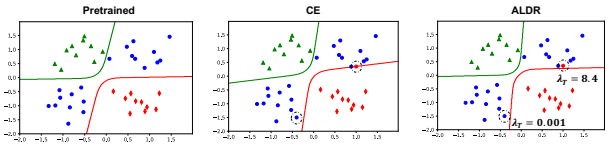


Figure 1. Left: pretrained model on a synthetic data; Middle: a new model learned by optimizing the CE loss after adding two examples in dashed circle; Right: a new model learned by optimizing the ALDR-KL loss after adding two examples in dashed circle. Shape denotes the true label, and color denotes the given label.

where $\alpha > 0$ is a hyper-parameter, and λ_0 is a relatively large value. To intuitively understand this loss, we consider when the given label is clean and noisy. If the given label is clean, then it is better to push the largest $f_k(\mathbf{x}), k \neq y$ to be small, hence a spiked $\mathbf{p}$ would be better, as a result $\text{KL}(\mathbf{p}, 1/K)$ would be large; so by maximizing λ it will push λ to be smaller. On the other hand, if the given label is noisy, then it is better to use uniform $\mathbf{p}$ (otherwise it may wrongly penalize more on predictions for the "good" class labels), as a result $\text{KL}(\mathbf{p}, 1/K)$ would be small, then by maximizing over λ it will push λ to be close to the prior λ_0.

To elaborate this intuition, we conduct an experiment on a simple synthetic data. We generate data for three classes from four normal distributions with means at (0.8,0.8), (0.8,-0.8), (-0.8,0.8), (-0.8,-0.8) and standard deviations equal to (0.3, 0.3) as shown in Figure 1 (left). To demonstrate the adpative λ mechanism of ALDR-KL, we first pretrain the model by optimizing the CE loss by 1000 epochs with momentum optimizer and learning rate as 0.01. Then, we add an extra hard but clean example from the $\circ$ class (the blue point in the dashed circle), and a noisy data from the $\circ$ class but mislabeled as the $\diamond$ class (the red point in the dashed circle), and then finetune the model with extra 100 epochs by optimizing our ALDR-KL loss with Algorithm 1 and optimizing CE loss with momentum SGD. We set the $\lambda_0 = 10, \alpha = 0.05$ for ALDR-KL loss. The learned model of using ALDR-KL loss (right) is more robust than that of using CE loss (middle). We also show the corresponding learned λ_T values at the last iteration for the two added examples, which show that the noisy data has a large $\lambda = 8.4$ and the clean data has a small $\lambda = 0.001$. We also plot the averaged λ_t and the KL-divergence values $\text{KL}(\mathbf{p}_t, 1/K)$ during the training process in Figure 3 of Appendix, which clearly shows the negative relationship between them. Additionally, we provide more synthetic experiments for LDR-KL loss by training from scratch in Appendix L.6, where the observations are consistent with the conclusion here.

The theoretical properties of ALDR-KL are stated below.

Theorem 5. *The ALDR-KL loss is All-k consistent when* $c_y = 0, \forall y, \lambda_0 \in (0, \infty), \alpha > \frac{\log K}{\lambda_0}$ *or* $\lambda_0 = \infty, \mathcal{C} = \{\mathbf{f} \in \mathbb{R}^K, \|\mathbf{f}\|_2 \leq B\}$, *and is Symmetric when* $\lambda_0 = \infty$.

Optimization for ALDR-KL loss: Using ALDR-KL loss and with a finite set of examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the empirical risk minimization problem becomes:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^n \psi^{\text{KL}}_{\alpha,\lambda_0}(\mathbf{x}_i, y_i), \tag{5}$$

where $\mathbf{w}$ denotes the parameter of the prediction function $f(\cdot; \mathbf{w})$. Below, we will present how to solve this problem without causing significant more burden than optimizing traditional CE loss. Recall that the formulation for LDR-KL in (3), it is easy to show that:

$$\psi^{\text{KL}}_{\alpha,\lambda_0}(\mathbf{x}, y) = \max_{\lambda \in \mathbb{R}_+} -\frac{\alpha}{2}(\lambda - \lambda_0)^2 + \lambda \log\left(\sum_{k=1}^K \exp\left(\frac{f_k(\mathbf{x}) + c_{k,y} - f_y(\mathbf{x})}{\lambda}\right)\right),$$

where we have used the closed-form solution for $\mathbf{p}$ given λ:

$$p^*_k(\lambda) \propto \exp\left(\frac{f_k(\mathbf{x}; \mathbf{w}) + c_{k,y} - f_y(\mathbf{x}; \mathbf{w})}{\lambda}\right), \forall k. \tag{6}$$

To compute the λ, first by optimality condition ignoring the non-negative constraint we have $-\text{KL}(\mathbf{p}^*, \frac{1}{K}) - \alpha(\lambda - \lambda_0) = 0$. This motivate us to update λ alternatively with $\mathbf{p}$. Given λ^i_{t-1} for an individual data $\mathbf{x}_i$ at the $(t-1)$-th iteration, we compute $\mathbf{p}^i_{t-1}$ according to (6) then we update λ^i_t according to the above equation with $\mathbf{p}^*$ replaced by $\mathbf{p}^i_{t-1}$ and a projection onto non-negative orthant $\mathbb{R}_+$:

$$\lambda^i_t = \left[\lambda_0 - \frac{1}{\alpha}\text{KL}(\mathbf{p}^i_{t-1}, \frac{1}{K})\right]_+. \tag{7}$$

For updating the model parameter $\mathbf{w}$, at iteration t suppose we have the current estimation for λ_t, we can compute the stochastic gradient estimation for $\mathbf{w}_{t-1}$ as:

$$G(\lambda^i_t, \mathbf{w}_{t-1}, \mathbf{x}_i) = \tag{8}$$
$$\nabla_{\mathbf{w}} \lambda^i_t \log \sum_{k=1}^K \exp\left(\frac{f_k(\mathbf{x}_i, \mathbf{w}_{t-1}) + c_{k,y} - f_y(\mathbf{x}_i, \mathbf{w}_{t-1})}{\lambda^i_t}\right)$$

Finally, we provide the algorithm using the momentum update for optimizing ALDR-KL loss in Alg. 1. It is notable that other updates (e.g., SGD, Adam) can be used for updating $\mathbf{w}_t$ as well. It is worth noting that we could also use gradient method to optimize λ and $\mathbf{p}$, but it will involve

Algorithm 1 Stochastic Optimization for ALDR-KL loss

Require: $\eta, \alpha, \beta, \lambda_0$
Random initialize model parameters $\mathbf{w}_0$; initialize λ^i_0 as λ_0 for each data sample, $i \in [n]$; initialize $\mathbf{m}_0 = 0$.
for $t = 1, ...T$ **do**
  Sample mini-batch of data indices $\mathcal{B}_t$
  for each sample $i \in \mathcal{B}_t$ **do**
    Compute $\mathbf{p}^i_{t-1}$ by E.q. 6 with λ^i_{t-1}, $f(\mathbf{x}_i, \mathbf{w}_{t-1})$.
    Update $\lambda^i_t = [\lambda_0 - \frac{1}{\alpha}\text{KL}(\mathbf{p}^i_{t-1}, \frac{1}{K})]_+$
    Compute $G(\lambda^i_t, \mathbf{w}_{t-1}, \mathbf{x}_i)$ by E.q. 8
  end for
  Compute mini-batch gradient estimation by: $G(\mathbf{w}_{t-1}) = \frac{1}{|\mathcal{B}_t|}\sum_{i \in \mathcal{B}_t} G(\lambda^i_t, \mathbf{w}_{t-1}, \mathbf{x}_i)$
  Compute $\mathbf{m}_t = \beta \mathbf{m}_{t-1} + (1-\beta) G(\mathbf{w}_{t-1})$
  Update $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \mathbf{m}_t$
end for

tuning two extra learning rates. Empirically we verify that Alg. 1 converges well for optimizing the ALDR-KL loss, which is presented in Appendix L for interested readers and the convergence analysis is left as a future work.

4. Experiments

In this section, we present the datasets, experimental settings, experimental results, and ablation study regarding the adaptiveness in terms of λ for ALDR-KL loss. For all the experiments unless specified otherwise, we manually add label noises to the training and validation data, but keep the testing data clean. A similar setup was used in (Zhang & Sabuncu, 2018), which simulates the noisy datasets in practice. We compare with all loss functions in Table 4 except for NCE, which is replaced by NCE+RCE following (Ma et al., 2020) and NCE+AGCE, NCE+AUL following (Zhou et al., 2021). Detailed expressions of these losses with hyper-parameters are shown in Table 4. We do not compare the Generalized JS Loss (GJS) because it relies on extra unsupervised data augmentation and need more computational cost (Englesson & Azizpour, 2021). For all experiments, we utilize the identical optimization procedures and models, only with different losses.

4.1. Benchmark Datasets with Synthetic Noise

We conduct experiments on 7 benchmark datasets, namely ALOI, News20, Letter, Vowel (Fan & Lin), Kuzushiji-49, CIFAR-100 and Tiny-ImageNet (Clanuwat et al., 2018; Deng et al., 2009). The number of classes vary from 11 to 1000. The statistics of the datasets are summarized in Table 5 in the Appendix. We manually add different noises to the datasets. Similar to previous works (Zhang & Sabuncu, 2018; Wang et al., 2019), we consider two different noise settings, uniform (symmetric) noise and class dependent (asymmetric) noise. For uniform noise, we impose a uniform random noise with the probability $\xi \in \{0.3, 0.6, 0.9\}$ such that a data label is changed to any label in $\{1, ..., K\}$ with a probability ξ. We would like to point out that the uniform noise rate 0.9 should be tolerable according the theory of a symmetric loss, which can tolerate uniform noise rate up to $(1-1/K) > 0.9$, where K is larger than 10 for the considered datasets. For class dependent noise, with the probability in $\xi \in \{0.1, 0.3, 0.5\}$, a data label could be changed to another label by following pre-defined transition rules. For letter dataset: $B \leftrightarrow D$, $C \leftrightarrow G$, $E \leftrightarrow F$, $H \leftrightarrow N$, $I \leftrightarrow L$, $K \leftrightarrow X$, $M \leftrightarrow W$, $O \leftrightarrow Q$, $P \leftrightarrow R$, $U \leftrightarrow V$. For news20 dataset: *comp.os.ms-windows.misc* $\leftrightarrow$ *comp.windows.x*, *comp.sys.ibm.pc.hardware* $\leftrightarrow$ *comp.sys.mac.hardware*, *rec.autos* $\leftrightarrow$ *rec.motorcycles*, *rec.sport.baseball* $\leftrightarrow$ *rec.sport.hockey*, *sci.crypt* $\leftrightarrow$ *sci.electronics*, *soc.religion.christian* $\leftrightarrow$ *talk.politics.misc*. For vowel dataset: $i \leftrightarrow I$, $E \leftrightarrow A$, $a :\leftrightarrow Y$, $C :\leftrightarrow O$, $u :\leftrightarrow U$. For ALOI, Kuzushiji-49, CIFAR-100 and Tiny-ImageNet datasets, we simulate class-dependent noise by flipping each class into the next one circularly with probability ξ.

For all datasets except for Kuzushiji-49, CIFAR-100 and Tiny-ImageNet, we uniformly randomly split 10% of the whole data as testing data, and the remaining as training set. For Kuzushiji-49 and CIFAR-100 dataset, we use its original train/test splitting. For Tiny-ImageNet, we use its original train/validation splitting and treat the validation data as the testing data in this work. For News20 dataset, because the original dimension is too large, we apply PCA to reduce the dimension to 80% variance level.

4.2. Experimental Settings

Since ALOI, News20, Letter and Vowel datasets are tabular data, we use a 2-layer feed forward neural network, with a number of neurons in the hidden layer same as the number of features or the number of classes (whichever is smaller) as the backbone model for these four datasets. ResNet18 is utilized as the backbone model for Kuzushiji-49, CIFAR100 and Tiny-ImageNet image datasets (He et al., 2016). We apply 5-fold-cross-validation to conduct the training and evaluation, and report the mean and standard deviation for the testing top-k accuracy, where $k \in \{1, 2, 3, 4, 5\}$. We use early stopping according to the highest accuracy on the validation data. For Kuzushiji-49 dataset, the top-k accuracy is first computed in class level, and then averaged across the classes according to (Clanuwat et al., 2018) because the classes are imbalanced.

We fix the weight decay as 5e-3, batch size as 64, and total running epochs as 100 for all the datasets except Kuzushiji, CIFAR100 and Tiny-ImageNet (we run 30 epochs for them because the data sizes are large). We utilize the momentum optimizer with the initial learning rate tuned in {1e-1, 1e-2, 1e-3} for all experiments. We decrease the learning rate by ten-fold at the end of the 50th and 75th epoch for 100-total-epoch experiments, and at the end of 10th and 20th epoch for 30-total-epoch experiments. For CE, MSE and MAE loss, there is no extra parameters need to be tuned; for CS and WW loss, we tune the margin parameter at {0.1, 1, 10}; for RLL loss, we tune the α parameter at {0.1, 1, 10}; for GCE and TGCE loss, we tune the power parameter q at {0.05, 0.7, 0.95} and set truncate parameter as 0.5 for TGCE; for SCE, as suggested in the original paper, we fix A=-4 and tune the simplified balance parameter between CE and RCE at {0.05, 0.5, 0.95}; for JS loss, we tune the balance parameter at {0.1, 0.5, 0.9}; for NCE+(RCE,AGCE,AUL) loss, we keep the default parameters for each individual loss, and tune the two weight parameters α/β in {0.1/9.9, 5/5, 9.9/0.1} following the original papers. For LDR-KL and ALDR-KL loss, we tune the DW regularization parameter λ or the λ_0 at $\{0.1, 1, 10\}$. We normalize model logits $\mathbf{f}(\mathbf{x}; \mathbf{w})$ by $\frac{\|\mathbf{f}(\mathbf{x};\mathbf{w})\|_1}{K}$ so that different λ values can achieve

Table 2. leaderboard for comparing 15 different loss functions on 7 datasets in 6 noise and 1 clean settings. The reported numbers are averaged ranks of performance. The smaller the better.

Loss Function	top-1	top-2	top-3	top-4	top-5	overall
ALDR-KL	**2.163**	**2.449**	**2.184**	**2.224**	**2.633**	**2.331**
LDR-KL	2.429	2.571	2.816	2.939	2.959	2.743
CE	4.714	4.592	4.224	4.449	4.388	4.473
SCE	4.898	4.776	4.429	4.837	4.327	4.653
GCE	5.816	5.347	6.041	5.571	5.551	5.665
TGCE	6.204	6.224	6.265	6.224	6.306	6.245
WW	7.673	6.592	6.082	5.959	5.551	6.371
JS	7.816	7.837	7.857	7.98	8.245	7.947
CS	8.673	8.878	8.816	8.592	8.735	8.739
RLL	10.224	10.204	10.224	10.49	10.49	10.326
NCE+RCE	9.694	10.633	10.878	10.612	10.714	10.506
NCE+AUL	10.449	10.959	11.102	11.224	11.122	10.971
NCE+AGCE	11.714	11.837	11.837	11.51	11.714	11.722
MSE	12.796	12.653	12.776	12.959	12.898	12.816
MAE	14.735	14.449	14.469	14.429	14.367	14.49

Table 3. Top-1 Accuracy on ILSVRC12 validation data with models trained on mini-webvision.

Loss	NCE+RCE	SCE	GCE	JS	CE	AGCE	LDR-KL	ALDR-KL
ACC	60.24	62.44	62.76	65.00	66.40	67.52	69.00	**69.92**

the different intended effects for LDR-KL and ALDR-KL loss. The margin parameter is fixed as 0.1 for all experiments. For ALDR-KL, we set the $\alpha = \frac{2\log K}{\lambda_0}$, so that $\lambda_t^i = \lambda_0 - \frac{1}{\alpha}\mathrm{KL}(\mathbf{p}_{t-1}(\mathbf{x}_i)\|\frac{1}{K}) \in [\lambda_0/2, \lambda_0]$, which maintains an appropriate adaptive adjustment level from λ_0.

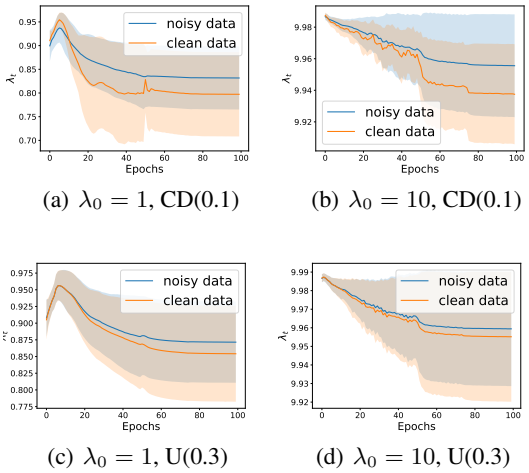
4.3. Leaderboard on seven benchmark data

We compare 15 losses on 7 benchmark datasets over 6 noisy and 1 clean settings for 5 metrics with 3,675 numbers. However, due to limit of space it is difficult to include the results here. Instead, we focus on the overall results on all datasets in all settings by presenting a leaderboard in Table 2. For each data in each setting, we rank different losses from 1 to 15 according to their performance from high to low (the smaller the rank value meaning the better of the performance). We average the ranks across all the benchmark datasets in all settings for top-k accuracy with different k and we also provide an overall rank.

We present the full results and their analysis for different performance metrics in the Appendix L. From the leaderboard in Table 2, we observe that the ALDR-KL loss is the strongest and followed by LDR-KL loss. Besides, the CE loss and two variants (SCE, GCE) are more competitive than the other loss functions. The symmetric MAE loss is non-surprisingly the worst which implies consistency is also an important factor for learning with noisy data.

4.4. Real-world Noisy Data

We evaluate the proposed methods on a real-world noisy dataset, namely webvision (Li et al., 2017). It contains more than 2.4 million of web images crawled from the Internet by using queries generated from the same 1,000 semantic concepts as the benchmark ILSVRC 2012 dataset. We mainly follow the settings on the previously work (Zhou et al., 2021; Ma et al., 2020; Englesson & Azizpour, 2021) by using a mini version, i.e., taking the first 50 classes from webvision as training data, and evaluating the results on ILSVRC12 validation data. More details about the experimental settings are summarized in the Appendix L.3. ResNet50 is utilized as the backbone model (He et al., 2016). The results summarized in Table 3 show ALDR-KL performs the best followed by LDR-KL. We notice that (Englesson & Azizpour, 2021) reports a higher result for JS than that in Table 3, which is due to different setups, especially with more epochs, larger weight decay and tuning of learning rates. Following the same experimental setup as (Englesson & Azizpour, 2021), our ALDR-KL and LDR-KL achieve 70.80 and 70.64 for top-1 accuracy, respectively, while JS's is reported as 70.36.



(a) $\lambda_0 = 1$, CD(0.1) (b) $\lambda_0 = 10$, CD(0.1)

(c) $\lambda_0 = 1$, U(0.3) (d) $\lambda_0 = 10$, U(0.3)

Figure 2. Averaged λ_t values for noisy samples and clean samples with error bar for ALDR-KL loss on Vowel dataset, 'U' is short for uniform noise, 'CD' is short for class-dependent noise.

4.5. Adaptive λ for ALDR-KL loss

Lastly, we conduct experiments to illustrate the adaptive λ_t^i values learned by using ALDR-KL loss on the Vowel dataset with uniform noise ($\xi = 0.3$) and class dependent noise ($\xi = 0.1$). We maintain the adaptive λ_t^i values for noisy data samples and clean data samples separately at each epoch. Then we output the mean and standard deviation for the adaptive λ_t^i values for noisy data and clean data. $\lambda_0 \in \{1, 10\}$ are adopted for ALDR-KL at this subsection. All the settings are identical with previous experimental setups except initial learning rate is fixed as $1e-2$ for this specific justification experiment. The results are presented in Figure 2. From the figures, we observe that ALDR-KL applies larger DW regularization parameter λ for the noisy data, which makes ALDR-KL loss closer to symmetric loss function, and therefore it is more robust to certain types of noises (e.g. uniform noise, simple non-uniform noise and class dependent noise) (Ghosh et al., 2017). Besides, the adaptive λ_t^i values are generally decreasing for both noisy

and clean data because the model is more certain about its predictions with more training steps.

4.6. Running Time

We provide the running time analysis for the LDR-KL and ALDR-KL loss in Table 17 in Appendix L.7. We observe that the running time of optimizing ALDR-KL is slightly higher than that of optimizing the CE loss, it is still comparable with optimizing other robust losses such as TGCE, SCE, RLL, etc.

5. Conclusions

We proposed a novel family of Label Distributional Robust (LDR) losses. We studied consistency and robustness of proposed losses. We also proposed an adaptive LDR loss to adapt to the noisy degree of individual data. Extensive experiments demonstrate the effectiveness of our new losses, LDR-KL and ALDR-KL losses.

6. Acknowledgement

This work is partially supported by NSF Career Award 2246753, NSF Grant 2246757 and NSF Grant 2246756. The work of Yiming Ying is partially supported by NSF (DMS-2110836, IIS-2110546, and IIS-2103450).

References

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and van den Oord, A. Are we done with imagenet? *CoRR*, abs/2006.07159, 2020. URL https://arxiv.org/abs/2006.07159.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 1567–1578, 2019.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607, 2020.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, March 2002. ISSN 1532-4435.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Duchi, C. J., Glynn, W. P., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2016.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Englesson, E. and Azizpour, H. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021.

Fan, R.-E. and Lin, C.-J. Libsvm data. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

Geusebroek, J.-M., Burghouts, G. J., and Smeulders, A. W. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

Ghosh, A., Kumar, H., and Sastry, P. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Goldberger, J. and Ben-Reuven, E. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H12GRgcxg.

Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I. W., Zhang, Y., and Sugiyama, M. Masking: A new perspective of noisy supervision. *CoRR*, abs/1805.08193, 2018. URL http://arxiv.org/abs/1805.08193.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. In Bengio, S., Wallach, H.,

Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/ad554d8c3b06d6b97ee76a2448bd7913-Paper.pdf.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531v1.

Howard, J. and Ruder, S. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146, 2018. URL http://arxiv.org/abs/1801.06146.

Hsu, C. W. and Lin, C. J. A comparision of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.

Janocha, K. and Czarnecki, W. M. On loss functions for deep neural networks in classification. *CoRR*, abs/1702.05659, 2017. URL http://arxiv.org/abs/1702.05659.

Kakade, S. M. and Shalev-Shwartz, S. On the duality of strong convexity and strong smoothness : Learning applications and matrix regularization. 2009.

Kornblith, S., Chen, T., Lee, H., and Norouzi, M. Why do better loss functions lead to less transferable features? In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=8twKpG5s8Qh.

Krizhevsky, A., Nair, V., and Hinton, G. CIFAR-10 and CIFAR-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6:1, 2009.

Lang, K. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pp. 331–339. Elsevier, 1995.

Lapin, M., Hein, M., and Schiele, B. Top-k multiclass svm. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28, pp. 325–333. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf.

Lapin, M., Hein, M., and Schiele, B. Loss functions for top-k error: Analysis and insights. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 1468–1477. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.163. URL https://doi.org/10.1109/CVPR.2016.163.

Lapin, M., Hein, M., and Schiele, B. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(7):1533–1554, 2018. doi: 10.1109/TPAMI.2017.2751607. URL https://doi.org/10.1109/TPAMI.2017.2751607.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Li, W., Wang, L., Li, W., Agustsson, E., and Gool, L. V. Webvision database: Visual learning and understanding from web data. *CoRR*, abs/1708.02862, 2017. URL http://arxiv.org/abs/1708.02862.

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., and Bailey, J. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pp. 6543–6553. PMLR, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pp. 2971–2980, 2017.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103:127–152, 2005.

Patel, D. and Sastry, P. Memorization in deep neural networks: Does the loss function matter? In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 131–142. Springer, 2021.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2233–2241. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.240. URL https://doi.org/10.1109/CVPR.2017.240.

Qi, Q., Guo, Z., Xu, Y., Jin, R., and Yang, T. A practical online method for distributionally deep robust optimization. *arXiv preprint arXiv:2006.10138*, 2020.

Shalev-Shwartz, S. and Wexler, Y. Minimizing the maximal loss: How and why. In *ICML*, pp. 793–801, 2016.

Tang, Y. Deep learning using support vector machines. *CoRR*, abs/1306.0239, 2013. URL http://arxiv.org/abs/1306.0239.

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(May):1007–1025, 2007.

Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. From imagenet to image classification: Contextualizing progress on benchmarks. In *ArXiv preprint arXiv:2005.11295*, 2020.

van Rooyen, B., Menon, A., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/45c48cce2e2d7fbdea1afc51c7c6ad26-Paper.pdf.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019.

Weston, J. and Watkins, C. Multi-class support vector machines. Technical report, Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May, 1998.

Yang, F. and Koyejo, S. On the consistency of top-k surrogate losses. In *International Conference on Machine Learning*, pp. 10727–10735. PMLR, 2020.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*, 5: 1225–1251, December 2004. ISSN 1532-4435.

Zhang, Y., Niu, G., and Sugiyama, M. Learning noise transition matrix from only noisy labels via total variation regularization. In *International Conference on Machine Learning*, pp. 12501–12512. PMLR, 2021.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Zhou, X., Liu, X., Jiang, J., Gao, X., and Ji, X. Asymmetric loss functions for learning with noisy labels. In *International conference on machine learning*, pp. 12846–12856. PMLR, 2021.

A. Explicit expressions of all relevant losses

Table 4. Comparison between different loss functions. Notice that A from SCE loss is a negative value, which is suggested to be set as -4 in the original paper (Wang et al., 2019). For JS loss, $\mathbf{m} = \pi_1 \mathbf{e}_y + (1-\pi_1)\mathbf{p}$. NA means not available or unknown. * means that the results are derived by us (proofs are included in the appendix J).

Category	Loss	$p_y = \frac{\exp(f_y)}{\sum_k \exp(f_k)}$, $\mathbf{p} = (p_1, \ldots, p_K)$, $\mathbf{e}_y$ one-hot vector	Top-1 consistent	All-k consistent	Symmetric	instance addaptivity
Traditional	CE	$\psi_{\text{CE}}(\mathbf{f}, y) = -\log(p_y)$	Yes	Yes	No	No
	CS (Crammer & Singer, 2002)	$\psi_{\text{CS}}(\mathbf{f}, y, c) = \max(0, \max_{k \neq y} f_k - f_y + c)$	No	No	No	No
	WW (Weston & Watkins, 1998)	$\psi_{\text{WW}}(\mathbf{f}, y, c) = \sum_{k \neq y} \max(0, f_k - f_y + c)$	No	No	No	No
Symmetric	MAE (RCE, clipped) (Ghosh et al., 2017)	$\psi_{\text{MAE}}(\mathbf{f}, y) = 2(1 - p_y)$	Yes	No	Yes	No
	NCE (Ma et al., 2020)	$\psi_{\text{NCE}}(\mathbf{f}, y) = \log(p_y) / \sum_k \log(p_k)$	Yes	No	Yes	No
	RLL (Patel & Sastry, 2021)	$\psi_{\text{RL}}(\mathbf{f}, y, \alpha) = -\log(\alpha + p_y) + \frac{1}{K-1} \sum_{i \neq y} \log(\alpha + p_i)$	No	No	Yes	No
Interpolation between CE and MAE	GCE (Zhang & Sabuncu, 2018)	$\psi_{\text{GCE}}(\mathbf{f}, y, q) = (1 - p_y^q)/q$ where $(q \in [0, 1])$	Yes	Yes* $(0 \leq q < 1)$	Yes $(q = 1)$	No
	SCE (Wang et al., 2019)	$\psi_{\text{SCE}}(\mathbf{f}, y, \alpha, A) = \alpha \psi_{\text{CE}}(\mathbf{f}, y) - (1-\alpha)\frac{A}{2} \psi_{\text{MAE}}(\mathbf{f}, y)$	Yes	Yes* $(0 \leq \alpha < 1)$	Yes $(\alpha = 0)$	No
	JS (Englesson & Azizpour, 2021)	$\psi_{\text{JS}}(\mathbf{f}, y, \pi_1) = (\pi_1 \text{KL}(\mathbf{e}_y, \mathbf{m}) + (1-\pi_1)\text{KL}(\mathbf{p}, \mathbf{m}))/Z$	NA	NA	Yes $(\pi_1 = 1)$	No
Bounded	MSE (Ghosh et al., 2017)	$\psi_{\text{MSE}}(\mathbf{f}, y) = 1 - 2p_y + \|\mathbf{p}\|_2^2$	Yes	Yes*	No	No
	TGCE (Zhang & Sabuncu, 2018)	$\psi_{\text{TGCE}}(\mathbf{f}, y, q, k) = (1 - \max(p_y, k)^q)/q$ where $(q \in [0, 1])$	NA	NA	No	No
Asymmetric	AGCE (Zhou et al., 2021)	$\psi_{\text{AGCE}}(\mathbf{f}, y, q) = ((a+1)^q - (a + p_y)^q)/q, (a > 0, q > 0)$	Yes	No	No	No
	AUL (Zhou et al., 2021)	$\psi_{\text{AUL}}(\mathbf{f}, y, q) = ((a - p_y)^q - (a-1)^q)/q \; (a > 1, q > 0)$	Yes	No	No	No
LDR	LDR-KL	$\psi_\lambda^{\text{KL}}(f, \mathbf{x}, y) = \lambda \log\left(\sum_{k=1}^K \exp\left(\frac{f_k + c_{y,k} - f_y}{\lambda}\right)\right)$	Yes* $(\lambda > 0)$		Yes* $(\lambda = \infty)$	No
	ALDR-KL	$\psi_{\alpha,\lambda_0}^{\text{KL}}(f, \mathbf{x}, y) = \max_{\lambda \geq 0} \psi_\lambda^{\text{KL}}(f, \mathbf{x}, y) - \frac{\alpha}{2}(\lambda - \lambda_0)^2$	Yes* $(\lambda_0 > 0, \alpha > \frac{\log K}{\lambda_0})$		Yes* $(\lambda_0 = \infty)$	Yes

B. Proof for LDR-KL Formulation and Special cases.

Proof. Let $\mathbf{q} = \mathbf{f} - \mathbf{f}_y + \mathbf{c}_y$, we want to solve:

$$\max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{q} - \lambda \sum_i \mathbf{p}_i \log K \mathbf{p}_i.$$

Using Langrangian duality theory to handle the constraint $\sum_i p_i = 1$, we have

$$\min_\eta \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{q} - \lambda \sum_i \mathbf{p}_i \log K \mathbf{p}_i - \eta(\sum_i p_i - 1).$$

By maximizing over $\mathbf{p}$, we have $p_i^* = \exp((q_i - \eta)/\lambda)$. With the Karush–Kuhn–Tucker (KKT) condition $\sum_i p_i^* = 1$, we have $p_i^* = \frac{\exp(q_i/\lambda)}{\sum_i \exp(q_i/\lambda)}$. By plugging this back we obtain the formulation of LDR-KL loss, i.e,

$$\psi_\lambda^{\text{KL}}(\mathbf{x}, y) = \lambda \log\left(\frac{1}{K} \sum_{k=1}^K \exp\left(\frac{f_k + c_{k,y} - f_y}{\lambda}\right)\right).$$

Extreme Case $\lambda = 0$:, with $\lambda = 0$, the optimal $\mathbf{p}_*$ will be one-hot vector that is only 1 for the k such that $f_k(\mathbf{x}) - f_y(\mathbf{x}) + c_{k,y}$ is largest. Since $f_y(\mathbf{x}) - f_y(\mathbf{x}) + c_{y,y} = 0$. Hence, we have $\sum_k p_k^*(f_k(\mathbf{x}) - f_y(\mathbf{x}) + c_{k,y}) = \max(0, \max_{k \neq y} f_k(\mathbf{x}) - f_y(\mathbf{x}) + c_y)$.

Extreme Case $\lambda = \infty$: , with $\lambda = \infty$, the optimal $\mathbf{p}_* = 1/K$. Hence, we have $\psi_\lambda^{\text{KL}}(\mathbf{x}, y) = \sum_k p_k^*(f_k(\mathbf{x}) - f_y(\mathbf{x}) + c_{k,y}) - \lambda \sum_i p_i^* \log p_i^* K = \frac{1}{K} \sum_{k=1}^K (f_k(\mathbf{x}) - f_y(\mathbf{x}) + c_{k,y})$.

□

C. Proof for Lemma 1

Proof. Fix $\mathbf{x}$, with $\mathbf{q} = (\Pr(y = 1|\mathbf{x}), \ldots, \Pr(y = K|\mathbf{x}))$, we will prove that if $\mathbf{f}$ is rank consistent with respect to $\mathbf{q}$, then $P_k(\mathbf{f}, \mathbf{q})$, i.e., $\mathbf{f}$ is top-k preserving with respect to $\mathbf{q}$ for any $k \in [K]$. Fix k, let $r_k(\mathbf{f})$ be any top-k selector of $\mathbf{f}$. First we have $1 - \sum_{m \in r_k(\mathbf{f})} \mathbf{q}_m \geq 1 - \sum_{m=1}^j \mathbf{q}_{[k]}$. If f is rank consistent, then the equality holds, i.e., $\sum_{m \in r_k(\mathbf{f})} q_m = \sum_{m=1}^j \mathbf{q}_{[k]}$. This is because that if the equality does not hold, then there exists $i \in r_k(f)$ and $j \in [K] \backslash r_k(f)$ (i.e., $f_j \leq f_i$) such that $q_j > q_i$. This would not happen as $\mathbf{f}$ is rank consistent with $\mathbf{q}$ according to assumption.

The following argument follows (Yang & Koyejo, 2020). If $\neg P_k(\mathbf{f}, \mathbf{q})$, then there exists $i \in [K]$ such that $q_i > q_{[k+1]}$ but $f_i \leq f_{[k+1]}$, or $q_i < q_{[k]}$ but $f_i \geq f_{[k]}$. In the first case, there is an r_k such that $i \notin r_k(\mathbf{f})$, because there are at least k indices $j \in [K]$, $j \neq i$ such that $f_j > f_i$. In the second case, there is an r_k such that $i \in r_k(\mathbf{f})$, because f_i is one of the top k values of $\mathbf{f}$. In either case, there is an r_k such that $\sum_{m \in r_k(\mathbf{f})} \mathbf{q}_m < \sum_{m=1}^j \mathbf{q}_{[k]}$. This contradicts to the conclusion derived before. Hence, we can conclude that if f is rank consistent with $\mathbf{q}$, we have $P_k(\mathbf{f}, q)$ for any $k \in [K]$. As a result, we conclude that $\inf_{\mathbf{f} \in \mathbb{R}^K : \neg P_k(\mathbf{f}, \mathbf{q})} L_\psi(\mathbf{f}, \mathbf{q}) > L_\psi(\mathbf{f}^*, \mathbf{q})$ if $\mathbf{f}^*$ is rank consistent with repect to $\mathbf{q}$. □

D. Proof for Theorem 1

Proof. For simplicity, we consider $\lambda = 1$. The extension to $\lambda > 0$ is trivial. Let us consider the optimization problem:

$$\min_{\mathbf{f}\in\mathbb{R}^K} F(\mathbf{f}) := \sum_k q_k \psi_k^\lambda(\mathbf{f}) = \sum_k q_k \log\left(\sum_l \exp(f_l - f_k + c(l,k))\right)$$

By the first-order optimality condition $\partial F(f)/\partial f_j = 0$, we have:

$$-q_j \frac{\sum_{l\neq j}\exp(f_l - f_j + c(l,j))}{\sum_l \exp(f_l - f_j + c(l,j))} + \sum_{k\neq j} q_k \frac{\exp(f_j - f_k + c(j,k))}{\sum_l \exp(f_l - f_k + c(l,k))} = 0$$

Move the negative term to the right and add $\frac{q_j}{\sum_l \exp(f_l - f_j + c(l,j))}$ to both sides:

$$\begin{aligned}
&q_j \frac{\sum_l \exp(f_l - f_j + c(l,j))}{\sum_l \exp(f_l - f_j + c(l,j))} \\
&= \sum_{k\neq j} \frac{q_k \exp(f_j - f_k + c(j,k))}{\sum_l \exp(f_l - f_k + c(l,k))} + \frac{q_j}{\sum_l \exp(f_l - f_j + c(l,j))} \\
&= \sum_k \frac{q_k \exp(f_j - f_k + c(j,k))}{\sum_l \exp(f_l - f_k + c(l,k))} \\
&= \exp(f_j) \sum_k \frac{q_k}{\sum_l \exp(f_l + c(l,k) - c(j,k))}
\end{aligned}$$

Moreover:

$$\begin{aligned}
&\sum_k \frac{q_k}{\sum_l \exp(f_l + c(l,k) - c(j,k))} \\
&= \frac{q_j}{\exp(f_j) + \sum_{l\neq j}\exp(f_l + c_j)} + \sum_{k\neq j} \frac{q_k}{\exp(f_k - c_k) + \sum_{l\neq k}\exp(f_l)} \\
&= \frac{q_j}{\exp(f_j) + \sum_{l\neq j}\exp(f_l + c_j)} + \sum_{k\neq j} \frac{q_k \exp(c_k)}{\exp(f_k) + \sum_{l\neq k}\exp(f_l + c_k)} \\
&= \sum_k \frac{q_k \exp(c(j,k))}{\exp(f_k) + \sum_{l\neq k}\exp(f_l + c_k)} \\
&= \frac{\sum_k q_k \exp(c(j,k))}{Z}
\end{aligned}$$

where $Z > 0$ is independent of j. Then

$$\exp(f_j^*) = \frac{q_j Z}{\sum_k q_k \exp(c(j,k))} = \frac{q_j Z}{q_j + \sum_{k\neq j} q_k \exp(c_k)}$$

If $c_k = c$, we have

$$\begin{aligned}
\exp(f_j^*) &= \frac{q_j Z}{\sum_k q_k \exp(c(j,k))} \\
&= \frac{q_j Z}{q_j + \sum_k q_k \exp(c_k) - q_j \exp(c_j)} \\
&= \frac{q_j Z}{\sum_k q_k \exp(c_k) - q_j(\exp(c_j) - 1)}
\end{aligned}$$

Hence the larger q_j, the larger f_j^*. For $q_i < q_j$, we have

$$\begin{aligned}
&1/\exp(f_i^*) - 1/\exp(f_j^*) \\
&= \frac{\exp(c)}{q_i Z} - \frac{\exp(c) - 1}{Z} - \frac{\exp(c)}{q_j Z} + \frac{\exp(c) - 1}{Z} > 0
\end{aligned}$$

It implies that $f_i^* < f_j^*$. □

E. Proof of Theorem 2

Definition 5. *For a function $R : \Omega \to \mathbb{R} \cap \{-\infty, \infty\}$ taking values on extended real number line, its convex conjugate is defined as*

$$R^*(\mathbf{u}) = \max_{\mathbf{p}\in\Omega} \mathbf{p}^\top \mathbf{u} - R(\mathbf{p})$$

Lemma 2. *If $R(\cdot)$ and Ω are (input-) symmetric, then R^* is also input-symmetric.*

Proof.

$$R^*(\mathbf{u}) = \max_{\mathbf{p}\in\Omega} \mathbf{p}^\top \mathbf{u} - R(\mathbf{p})$$

For any i, j, let $\widehat{\mathbf{u}}$ be a shuffled version of $\mathbf{u}$ such that $\widehat{\mathbf{u}}_i = \mathbf{u}_j, \widehat{\mathbf{u}}_j = \mathbf{u}_i, \widehat{\mathbf{u}}_k = \mathbf{u}_k, \forall k \neq i, j$. Then

$$\begin{aligned}
R^*(\widehat{\mathbf{u}}) &= \max_{\mathbf{p}\in\Omega} \mathbf{p}^\top \widehat{\mathbf{u}} - R(\mathbf{p}) = \max_{\mathbf{p}\in\Omega} \sum_k p_k \widehat{\mathbf{u}}_k - R(\mathbf{p}) \\
&= \max_{\mathbf{p}\in\Omega} p_i \mathbf{u}_j + p_j \mathbf{u}_i + \sum_{k\neq i,j} p_k \mathbf{u}_k - R(\mathbf{p}) \\
&= \max_{\mathbf{p}\in\Omega} \widehat{\mathbf{p}}^\top \mathbf{u} - R(\mathbf{p}) = \max_{\widehat{\mathbf{p}}\in\Omega} \widehat{\mathbf{p}}^\top \mathbf{u} - R(\widehat{\mathbf{p}}) \\
&= R^*(\mathbf{u})
\end{aligned}$$

where $\widehat{\mathbf{p}}$ is a shuffled version of $\mathbf{p}$. □

Proof. **Proof of Theorem 2** , without loss of generalization, we assume $\lambda = 1$.

$$\begin{aligned}
\psi_y(\mathbf{f}) &= \max_{\mathbf{p}\in\Omega} \sum_k p_k (f_k - f_y + c_{k,y}) - R(\mathbf{p}) \\
&= R^*(\mathbf{f} - f_y \mathbf{1} + \mathbf{c}_y)
\end{aligned}$$

where $\mathbf{c}_y = c(1 - \mathbf{e}_y)$, $\mathbf{e}_y$ is the y-th column of identity matrix. Let us consider the optimization problem:

$$\begin{aligned}
\mathbf{f}^* &= \arg\min_{\mathbf{f}\in\mathbb{R}^K} F(\mathbf{f}) := \sum_k q_k \psi_k(\mathbf{f}) \\
&= \sum_k q_k R^*(\mathbf{f} - f_k \mathbf{1} + \mathbf{c}_k)
\end{aligned}$$

Consider $\mathbf{q}$ such that $q_1 < q_2$, we prove that $\mathbf{f}^*$ is order preserving. We prove by contradiction. Assume $\mathbf{f}_1^* > \mathbf{f}_2^*$. Define $\widehat{\mathbf{f}}$ as $\widehat{\mathbf{f}}_1 = \mathbf{f}_2^*$ and $\widehat{\mathbf{f}}_2 = \mathbf{f}_1^*$ and $\widehat{\mathbf{f}}_k = \mathbf{f}_k^*, k > 2$. Then

$$\begin{aligned}
F(\widehat{\mathbf{f}}) &= \sum_k q_k R^*(\widehat{\mathbf{f}} - \widehat{f}_k \mathbf{1} + \mathbf{c}_k) \\
&= q_1 R^*(\widehat{\mathbf{f}} - \widehat{f}_1 \mathbf{1} + \mathbf{c}_1) + q_2 R^*(\widehat{\mathbf{f}} - \widehat{f}_2 \mathbf{1} + \mathbf{c}_2) + \sum_{k>2} q_k R^*(\widehat{\mathbf{f}} - \widehat{f}_k \mathbf{1} + \mathbf{c}_k) \\
&= q_1 R^*(\widehat{\mathbf{f}} - f_2^* \mathbf{1} + \mathbf{c}_1) + q_2 R^*(\widehat{\mathbf{f}} - f_1^* \mathbf{1} + \mathbf{c}_2) + \sum_{k>2} q_k R^*(\mathbf{f} - f_k^* \mathbf{1} + \mathbf{c}_k)
\end{aligned}$$

Then

$$F(\widehat{\mathbf{f}}) - F(\mathbf{f}^*) = q_1 (R^*(\widehat{\mathbf{f}} - f_2^* \mathbf{1} + \mathbf{c}_1) - R^*(\mathbf{f} - f_1^* \mathbf{1} + \mathbf{c}_1)) + q_2 (R^*(\widehat{\mathbf{f}} - f_1^* \mathbf{1} + \mathbf{c}_2) - R^*(\mathbf{f} - f_2^* \mathbf{1} + \mathbf{c}_2))$$

Since $[\widehat{\mathbf{f}} - f_2^* \mathbf{1} + \mathbf{c}_1]_1 = 0$, $[\widehat{\mathbf{f}} - f_2^* \mathbf{1} + \mathbf{c}_1]_2 = f_1^* - f_2^* + c$, $[\widehat{\mathbf{f}} - f_2^* \mathbf{1} + \mathbf{c}_1]_k = f_k^* - f_2^* + c, k > 2$, similarly $[\mathbf{f} - f_2^* \mathbf{1} + \mathbf{c}_2]_1 = f_1^* - f_2^* + c$, $[\mathbf{f} - f_2^* \mathbf{1} + \mathbf{c}_2]_2 = 0$, $[\mathbf{f}^* - f_2^* \mathbf{1} + \mathbf{c}_2]_k = f_k^* - f_2^* + c, k > 2$, hence $R^*(\widehat{\mathbf{f}} - f_2^* \mathbf{1} + \mathbf{c}_1) = R^*(\mathbf{f} - f_2^* \mathbf{1} + \mathbf{c}_2)$. Similarly $R^*(\mathbf{f}^* - f_1^* \mathbf{1} + \mathbf{c}_1) = R^*(\widehat{\mathbf{f}} - f_1^* \mathbf{1} + \mathbf{c}_2)$. Hence, we have

$$F(\widehat{\mathbf{f}}) - F(\mathbf{f}^*) = (q_2 - q_1)(R^*(\mathbf{f}^* - f_1^* \mathbf{1} + \mathbf{c}_1) - R^*(\mathbf{f}^* - f_2^* \mathbf{1} + \mathbf{c}_2))$$

Define $\mathbf{f}^\sharp = \mathbf{f}^* - f_1^* \mathbf{1} + \mathbf{c}_1$ with $[\mathbf{f}^\sharp]_1 = 0, [\mathbf{f}^\sharp]_k = f_k^* - f_1^* + c, k > 1$, and $\mathbf{f}^\dagger = \mathbf{f}^* - f_2^* \mathbf{1} + \mathbf{c}_2$ with $[\mathbf{f}^\dagger]_2 = 0, [\mathbf{f}^\dagger]_k = f_k^* - f_2^* + c, k \neq 2$. Let $\mathbf{f}^\ddagger$ be a shuffled version of $\mathbf{f}^\dagger$ with $[\mathbf{f}^\ddagger]_1 = 0, [\mathbf{f}^\ddagger]_k = f_1^* - f_2^* + c, k = 2, [\mathbf{f}^\ddagger]_k = f_k^* - f_2^* + c, k > 2$. Under the assumption that $f_1^* > f_2^*$, we have $\mathbf{f}^\sharp \leq \mathbf{f}^\ddagger$. Then

$$F(\widehat{\mathbf{f}}) - F(\mathbf{f}^*) = (q_2 - q_1)(R^*(\mathbf{f}^\sharp) - R^*(\mathbf{f}^\ddagger))$$

Note that

$$R^*(\mathbf{f}) = \max_{\mathbf{p}\in\Omega} \mathbf{p}^\top \mathbf{f} - R(\mathbf{p})$$

Below, we prove $R^*(\mathbf{f}^\sharp) - R^*(\mathbf{f}^\ddagger) < 0$. We will consider two scenarios. First, consider $f_2^* - f_1^* + c > 0$. Let $\mathbf{p}^* = \arg\max_{\mathbf{p}\in\Omega} \sum_{k>1} p_k [\mathbf{f}^\sharp]_k - R(\mathbf{p})$. We have $\mathbf{f}^\sharp \in \nabla R(\mathbf{p}^*)$. Since $[\mathbf{f}^\sharp]_2 > 0$, we have $p_2^* > 0$ (due to the assumption on R). Hence $R^*(\mathbf{f}^\sharp) = \max_{\mathbf{p}\in\Omega} \sum_{k>1} p_k [\mathbf{f}^\sharp]_k - R(\mathbf{p}) = \sum_{k>1} p_k^* [\mathbf{f}^\sharp]_k - R(\mathbf{p}^*) < \sum_{k>1} p_k^* [\mathbf{f}^\ddagger]_k - R(\mathbf{p}^*) \leq R^*(\mathbf{f}^\ddagger)$. Next, consider $f_2^* - f_1^* + c \leq 0$. By the convexity of R^*, we have $R^*(\mathbf{f}^\sharp) - R^*(\mathbf{f}^\dagger) \leq \nabla R^*(\mathbf{f}^\sharp)(\mathbf{f}^\sharp - \mathbf{f}^\dagger) = [\nabla R^*(\mathbf{f}^\sharp)]_1 (f_2 - f_1 - c) + [\nabla R^*(\mathbf{f}^\sharp)]_2 (f_2^* - f_1^* + c) + \sum_{k>2} [\nabla R^*(\mathbf{f}^\sharp)]_k (f_2 - f_1) \leq c([\nabla R^*(\mathbf{f}^\sharp)]_2 - [\nabla R^*(\mathbf{f}^\sharp)]_1) + \sum_k [\nabla R^*(\mathbf{f}^\sharp)]_k (f_2^* - f_1^*) < 0$ due to that (i) $[\nabla R^*(\mathbf{f}^\sharp)]_2 \leq [\nabla R^*(\mathbf{f}^\sharp)]_1$ due to $[\mathbf{f}^\sharp]_2 \leq [\mathbf{f}^\sharp]_1$; (ii) and $\sum_k [\nabla R^*(\mathbf{f}^\sharp)]_k = \sum_k p_k^* > 0$. This is because $\mathbf{p}^*$ cannot be all zeros otherwise increasing the first coordinate of $\mathbf{p}^*$ will decrease the value of $R(\mathbf{p})$ and therefore increase the value of Q. However, such a solution $\mathbf{p}^*$ is impossible since for the function $Q(\mathbf{p}) = \mathbf{p}^\top \mathbf{f}^\sharp - R(\mathbf{p})$, which has a unique maximizer due to strong concavity of $Q(\mathbf{p})$. This will prove that $R^*(\mathbf{f}^\sharp) - R^*(\mathbf{f}^\ddagger) < 0$. As a result, we prove that if $f_1^* > f_2^*$, then $F(\widehat{\mathbf{f}}) < F(\mathbf{f}^*)$, which is impossible since $\mathbf{f}^*$ is the minimizer of F. Hence, we have $f_1^* \leq f_2^*$.

Next, we prove that $f_1^* < f_2^*$. Assume $f_1^* = f_2^*$, we establish a contradiction. By the first-order optimality condition, we have

$$\begin{aligned}\frac{\partial F(f)}{\partial f_k} &= q_k \frac{\psi_k(\mathbf{f})}{\partial f_k} + \sum_{l\neq k} q_l \frac{\partial \psi_l(\mathbf{f})}{\partial f_k} \\ &= q_k (\mathbf{e}_k - 1)^\top \nabla R^*(\mathbf{f} - f_k \mathbf{1} - \mathbf{c}_k) + \sum_{l\neq k} q_l \mathbf{e}_k^\top \nabla R^*(\mathbf{f} - f_l \mathbf{1} - \mathbf{c}_l) = 0\end{aligned}$$

Hence

$$q_k \mathbf{1}^\top \nabla R^*(\mathbf{f}^* - f_k^* \mathbf{1} - \mathbf{c}_k) = \mathbf{e}_k^\top \sum_l q_l \nabla R^*(\mathbf{f}^* - f_l^* \mathbf{1} - \mathbf{c}_l)$$

Next, we prove that for any $\mathbf{f}$ such that $[\mathbf{f}]_1 = [\mathbf{f}]_2$, then $[\nabla R^*(\mathbf{f})]_1 = [\nabla R^*(\mathbf{f})]_2$. To this end, we consider the equation, $R(\mathbf{p}) + R^*(\mathbf{f}) = \mathbf{p}^\top \mathbf{f}$, where $\mathbf{p} = \arg\max_{\mathbf{p}\in\Omega} Q(\mathbf{p}) := \mathbf{p}^\top \mathbf{f} - R(\mathbf{p})$ satisfying $\mathbf{p} = \nabla R^*(\mathbf{f}) \in \Omega$. If $[\nabla R^*(\mathbf{f})]_1 \neq [\nabla R^*(\mathbf{f})]_2$, which means $p_1 \neq p_2$. We can define another vector $\mathbf{p}' = (p_2, p_1, p_3, \ldots, p_K) \in \Omega, \mathbf{p}' \neq \mathbf{p}$, which satisfies $Q(\mathbf{p}') = \mathbf{p}'^\top \mathbf{f} - R(\mathbf{p}') = Q(\mathbf{p})$, which contradicts to the fact that $Q(\cdot)$ has a unique maximizer. Then for any $\mathbf{f}$ with $[\mathbf{f}]_1 = [\mathbf{f}]_2$, we have $[\mathbf{f} - f_l \mathbf{1} - \mathbf{c}_l]_1 = [\mathbf{f} - f_l \mathbf{1} - \mathbf{c}_l]_2$. As a result,

$$q_1 \mathbf{1}^\top \nabla R^*(\mathbf{f}^* - f_1^* \mathbf{1} - \mathbf{c}_1) = q_2 \mathbf{1}^\top \nabla R^*(\mathbf{f}^* - f_2^* \mathbf{1} - \mathbf{c}_2)$$

It is clear that the above equation is impossible if $f_1^* = f_2^*$ since $\nabla R^*(\mathbf{f})$ is symmetric, $q_1 < q_2$ and $\mathbf{1}^\top \nabla R^*(\mathbf{f}) \neq 0, \forall \mathbf{f}$ with $[f]_1 = 0$. The later is due to the assumption $\nabla R(0) < 0$, the optimal solution $\mathbf{p} = \arg\max_{\mathbf{p}\in\Omega} Q(\mathbf{p}) := \mathbf{p}^\top \mathbf{f} - R(\mathbf{p})$ satisfying $\mathbf{p} = \nabla R^*(\mathbf{f}) \in \Omega$ cannot be all zeros (otherwise we can find a better solution by increasing the first component of $\mathbf{p}$ with a larger value of $\mathbf{p}^\top \mathbf{f} - R(\mathbf{p})$. □

F. Proof for Theorem 3

Let $\mathbf{q} = \mathbf{q}(\mathbf{x})$ be class distribution for a given $\mathbf{x}$. Let π_k^q denote the index of q whose corresponding value is the k-th largest entry. We define $\psi(\mathbf{f}, l) = \psi(f, \mathbf{x}, l)$.

Lemma 3. *If ψ is top-$1, 2$ consistent simultaneously when class number $K \geq 3$, then $\exists\, \mathbf{q} = \mathbf{q}(\mathbf{x})$ such that $q_{[3]} < q_{[2]} < q_{[1]}$ makes $\psi(\mathbf{f}^*, \pi_1^q) < \psi(\mathbf{f}^*, \pi_2^q) < \psi(\mathbf{f}^*, \pi_3^q)$, here $\mathbf{f}^* = \inf L_\psi(\mathbf{f}, \mathbf{q}) = \sum_l q_l \psi(\mathbf{f}, l)$.*

Proof. Without loss of generality, we consider $q_1 > q_2 > q_3$ and prove that if $\psi(\mathbf{f}^*, 1) < \psi(\mathbf{f}^*, 2) < \psi(\mathbf{f}^*, 3)$ does not hold, then we can construct $\mathbf{q}'$ such that $q'_{[3]} < q'_{[2]} < q'_{[1]}$ and $\psi(\mathbf{f}^*, \pi_1^{q'}) < \psi(\mathbf{f}^*, \pi_2^{q'}) < \psi(\mathbf{f}^*, \pi_3^{q'})$.

If $\psi(\mathbf{f}^*, 1) \geq \psi(\mathbf{f}^*, 2)$ and we construct $\mathbf{q}'$ such that $q'_{[3]} < q'_{[2]} < q'_{[1]}$ and $\psi(\hat{\mathbf{f}}^*, \pi_1^{q'}) < \psi(\hat{\mathbf{f}}^*, \pi_2^{q'})$, where $\hat{\mathbf{f}}^*$ be the optimal solution to $\inf L_\psi(\mathbf{f}, \mathbf{q}') = \sum_l q'_l \psi(\mathbf{f}, l)$. Then we construct $\mathbf{q}'$ by switching the 1st and 2nd entry of $\mathbf{q}$. Then we have $L_\psi(\mathbf{f}^*, \mathbf{q}) \geq L_\psi(\mathbf{f}^*, \mathbf{q}')$ because $\psi(\mathbf{f}^*, 1) q_1 + \psi(\mathbf{f}^*, 2) q_2 \geq \psi(\mathbf{f}^*, 1) q'_1 + \psi(\mathbf{f}^*, 2) q'_2$ due to $q'_1 = q_2, q'_2 = q_1$ and $q_1 > q_2$. Due to that ψ is top-1, 2 consistent, it is top-1,2 calibrated and hence $P_k(\mathbf{f}^*, \mathbf{q})$ and $P_k(\hat{\mathbf{f}}^*, \mathbf{q}')$ hold for $k = 1, 2$. As a result, we can prove that $\mathbf{f}_1^* > \mathbf{f}_2^* > \mathbf{f}_3^*$ and $\hat{\mathbf{f}}_2^* > \hat{\mathbf{f}}_1^* > \mathbf{f}_3^*$. As a result, $\mathbf{f}_* \neq \hat{\mathbf{f}}_*$. Next, we prove that $\psi(\hat{\mathbf{f}}^*, \pi_1^{q'}) < \psi(\hat{\mathbf{f}}^*, \pi_2^{q'})$. If $\psi(\hat{\mathbf{f}}^*, \pi_1^{q'}) \geq \psi(\hat{\mathbf{f}}^*, \pi_2^{q'})$, then $L_\psi(\hat{\mathbf{f}}^*, \mathbf{q}') \geq L_\psi(\hat{\mathbf{f}}^*, \mathbf{q})$ because $\psi(\hat{\mathbf{f}}^*, 1) q'_1 + \psi(\hat{\mathbf{f}}^*, 2) q'_2 \geq \psi(\hat{\mathbf{f}}^*, 1) q_1 + \psi(\hat{\mathbf{f}}^*, 2) q_2$ due to $\psi(\hat{\mathbf{f}}^*, 2) = \psi(\hat{\mathbf{f}}^*, \pi_1^{q'})$ and $\psi(\hat{\mathbf{f}}^*, 1) = \psi(\hat{\mathbf{f}}^*, \pi_2^{q'})$ and $q_1 > q_2$. Hence, we derive a fact that $L_\psi(\mathbf{f}^*, \mathbf{q}) \geq L_\psi(\mathbf{f}^*, \mathbf{q}') \geq L_\psi(\hat{\mathbf{f}}^*, \mathbf{q}') \geq L_\psi(\hat{\mathbf{f}}^*, \mathbf{q})$ under the assumption that $\psi(\hat{\mathbf{f}}^*, \pi_1^{q'}) \geq \psi(\hat{\mathbf{f}}^*, \pi_2^{q'})$, which is impossible as $\hat{\mathbf{f}}^*$ is not the optimal solution to $\inf L_\psi(\mathbf{f}, \mathbf{q}) = \sum_l q_l \psi(\mathbf{f}, l)$. Thus, we have $\psi(\hat{\mathbf{f}}^*, \pi_1^{q'}) < \psi(\hat{\mathbf{f}}^*, \pi_2^{q'})$. As a result, there exists $\mathbf{q}'$ such that $q'_{[3]} < q'_{[2]} < q'_{[1]}$ and $\psi(\mathbf{f}^*, \pi_1^{q'}) < \psi(\mathbf{f}^*, \pi_2^{q'})$.

If $\psi(\mathbf{f}^*, 3) \geq \psi(\mathbf{f}^*, 2)$ and construct $\mathbf{q}'$ such that $q'_{[3]} < q'_{[2]} < q'_{[1]}$ and $\psi(\hat{\mathbf{f}}^*, \pi_2^{q'}) < \psi(\hat{\mathbf{f}}^*, \pi_3^{q'})$. This case can be proved similarly as above.

□

Based on above lemma, we can finish the proof of Theorem 3. For any $f(\mathbf{x})$, let $\psi_f = (\psi(f, \mathbf{x}, 1), \ldots, \psi(f, \mathbf{x}, K))$. Then $\min_f \sum_l \psi(f, \mathbf{x}, l) q_l = \min_{\psi_f \in \mathbb{R}^K} \psi_f^\top \mathbf{q}$. Since $\psi_f^\top 1 = C$ due to the symmetry property. As a result, there exists $\psi_{f_*} \in \arg\min_{\psi_f \in \mathbb{R}^K} \psi_f^\top \mathbf{q}, s.t. \quad \psi_f^\top 1 = C, \psi_f \geq 0$ such that it is at a vertex of the feasible region (a generalized simplex), i.e., there only exist one $i \in [K]$ such that $[\psi_{f_*}]_i \neq 0$. This holds for any $\mathbf{q}$. Combining with the above lemma, we can conclude a contradiction, i.e., a non-negative symmetric loss cannot be top-1, 2 consistent simultaneously.

G. Proof for Theorem 4

Proof. When $\lambda = \infty$, the loss function becomes: $\psi_\infty^{\text{KL}}(\mathbf{x}, y) = \frac{1}{K} \sum_{k=1}^K (f_k(\mathbf{x}) - f_y(\mathbf{x}) + c(k, y))$. To show it is symmetric: $\sum_{j=1}^K \psi_\infty^{\text{KL}}\big(\mathbf{x}, y = j\big) = C, \ \forall f(\mathbf{x})$, we have

$$\sum_{j=1}^K \psi_\infty^{\text{KL}}\big(\mathbf{x}, y = j\big) = \frac{1}{K} \sum_{j=1}^K \sum_{k=1}^K (f_k(\mathbf{x}) - f_j(\mathbf{x}) + c(k, j)) = \frac{1}{K} \sum_{j=1}^K \sum_{k=1}^K c(k, j) = C$$

where C is a constant only depending on the predefined margins.

To prove All-k consistency, we prove that minimizing the expected loss function is rank preserving, i.e., $\mathbf{f}^* = \arg\min_{\mathbf{f} \in \mathcal{C}} \sum_k q_k \psi_k(\mathbf{f})$ is rank preserving, where $\psi_y(\mathbf{f}) = \psi_\infty^{\text{KL}}(\mathbf{x}, y) = \frac{1}{K} \sum_{k=1}^K ([\mathbf{f}]_k - [\mathbf{f}]_y + c(k, y))$. We can write $\psi_y(\mathbf{f}) = \frac{1}{K} \mathbf{a}_y^\top \mathbf{f} + \frac{(K-1)c_y}{K}$, where $[\mathbf{a}_y]_y = -1, [\mathbf{a}_y]_k = 1, \forall k \neq y$. As a result, we have $\sum_k q_k \psi_k(\mathbf{f}) = (\sum_k q_k \mathbf{a}_k)^\top \mathbf{f} + \frac{(K-1)\sum_k q_k c_k}{K}$, where $[\sum_k q_k \mathbf{a}_k]_i = -q_i + \sum_{j \neq i} q_j = 1 - 2q_i$. Then minimizing $\arg\min_{\|\mathbf{f}\|_2 \leq B} \sum_k q_k \psi_k(\mathbf{f}) = \arg\max_{\|\mathbf{f}\|_2 \leq B} \mathbf{f}^\top \hat{\mathbf{a}}$, where $[\hat{\mathbf{a}}]_i = 2q_i - 1$. As a result, $\mathbf{f}^* = \frac{B\hat{\mathbf{a}}}{\|\hat{\mathbf{a}}\|_2}$. Hence, if $q_i > q_j$, then $\mathbf{f}_i^* > \mathbf{f}_j^*$, which proves the rank preserving property of the optimal solution. Then, the conclusion follows the Lemma 1.

□

H. Proof for Theorem 5

Proof. We first consider $\alpha, \lambda_0 \in (0, \infty)$. For the ALDR-KL loss,

$$\begin{aligned}
\psi_{\alpha, \lambda_0}^{\text{KL}}(\mathbf{x}, y) = & \max_{\lambda \in \mathbb{R}_+} \max_{\mathbf{p} \in \Delta_K} \sum_{k=1}^K p_k (\delta_{k,y}(f(\mathbf{x})) + c_{k,y}) \\
& - \lambda \text{KL}(\mathbf{p}, \frac{1}{K}) - \frac{\alpha}{2}(\lambda - \lambda_0)^2 \\
= & \max_{\lambda \in \mathbb{R}_+} \lambda \log \frac{1}{K} \sum_{k=1}^K \exp(\frac{f_k - f_y + c(k, y)}{\lambda}) - \frac{\alpha}{2}(\lambda - \lambda_0)^2
\end{aligned} \tag{9}$$

Then the expected loss is given by

$$F(\mathbf{f}) = \sum_k q_k \psi_{\alpha, \lambda_0}^{\text{KL}}(\mathbf{x}, k) = \sum_k q_k \max_{\lambda \in \mathbb{R}_+} \lambda \log \frac{1}{K} \sum_{l=1}^K \exp(\frac{f_l - f_k + c(l, k)}{\lambda}) - \frac{\alpha}{2}(\lambda - \lambda_0)^2$$

$$= \sum_k q_k \max_{\lambda \in \mathbb{R}_+} (-f_k + \log \frac{1}{K} \sum_{l=1}^K \exp(\frac{f_l + c(l, k)}{\lambda}) - \frac{\alpha}{2}(\lambda - \lambda_0)^2).$$

Assuming that $c_{l,k} = 0$ (i.e., no margin is used), then we have

$$F(\mathbf{f}) = \sum_k q_k \psi_{\alpha, \lambda_0}^{\text{KL}}(\mathbf{x}, k) = \sum_k -q_k f_k + \sum_k q_k \max_{\lambda \in \mathbb{R}_+} \log \frac{1}{K} \sum_{l=1}^K \exp(\frac{f_l}{\lambda}) - \frac{\alpha}{2}(\lambda - \lambda_0)^2$$

$$= \sum_k -q_k f_k + \max_{\lambda \in \mathbb{R}_+} \log \frac{1}{K} \sum_{l=1}^K \exp(\frac{f_l}{\lambda}) - \frac{\alpha}{2}(\lambda - \lambda_0)^2$$

Similar to Theorem 1, we need to show the optimal solution $\mathbf{f}^*$ to the above problem is rank preserving as $\mathbf{q}$. Denote by

the optimal solution to λ as λ_* and to $\mathbf{p}$ as $\mathbf{p}_*$. Then $[\mathbf{p}_*]_j = \exp((f_j - f_y + c_{j,y})/\lambda_*)/\sum_k \exp((f_k - f_y + c_{k,y})/\lambda_*)$. Then $\partial F/\partial \mathbf{f}$ is same as $\partial F_{\lambda^*}/\partial \mathbf{f}$. Hence as long as $\lambda_* > 0$ is finite, we can follow the proof of Theorem 1 to prove the rank preserving. We know that $\lambda_* = [\lambda_0 - \frac{1}{\alpha}\mathrm{KL}(\mathbf{p}_*, 1/K)]_+$. Hence, if λ_0 is finite, $\mathrm{KL}(\mathbf{p}_*, 1/K)$ is clearly finite, so λ_* is finite. $\lambda_* > 0$ is ensured if $\frac{\log K}{\lambda_0} < \alpha$ as $\max_{\mathbf{p}\in\Delta} \mathrm{KL}(\mathbf{p}, 1/K) = \log K$. Then the conclusion follows.

If $\lambda_0 = \infty$, then we have $\lambda_* = \infty$; otherwise the loss is negative infinity. As a result, $\mathbf{p}_* = 1/K$. Then it reduces to that of Theorem 4.

□

I. Efficient Computation of LDR-k–KL loss.

As an interesting application of our general LDR-k loss, we restrict our attention to the use of KL divergence for the regularization term $R(\mathbf{p}) = \sum_{i=1} p_i \log(Kp_i)$ in the definition of E.q. 10,

$$\psi_\lambda^k(\mathbf{x}, y) = \max_{\mathbf{p}\in\Omega(k)} \sum_l p_l(\delta_{l,y}(f(\mathbf{x})) + c_{l,y}) - \lambda R(\mathbf{p}) \tag{10}$$

to which we refer as LDR-k-KL loss. In particular, we consider how to efficiently compute the LDR-k-KL loss. We present an efficient solution with $O(K \log K)$ complexity.

Theorem 6. *Consider the problem* $\max_{\mathbf{p}\in\Omega(k)} \sum_i p_i q_i - \lambda \sum_{i=1} p_i \log(Kp_i)$. *To compute the optimal solution, let the* $\mathbf{q}$ *to be sorted and* $[i]$ *denote the index for its* i*-th largest value. Given* $a \in [K]$ *define a vector* $\mathbf{p}(a)$ *as*

$$[\mathbf{p}(a)]_{[i]} = \frac{1}{k}, i < a,$$

$$[\mathbf{p}(a)]_{[i]} = \exp(\frac{\mathbf{q}_{[i]}}{\lambda} - 1) \times \min(\frac{1}{K}, \frac{1 - \frac{a-1}{k}}{\sum_{j=a}^K \exp(\frac{\mathbf{q}_{[j]}}{\lambda} - 1)}), \quad i \geq a$$

The optimal solution $\mathbf{p}_*$ *is given by* $\mathbf{p}(a)$ *such that* a *is the smallest number in* $\{1, \ldots, K\}$ *satisfying* $[\mathbf{p}(a)]_{[a]} \leq \frac{1}{k}$. *The overall time complexity is* $O(K \log(K))$.

Given that the algorithm would be simple with the presented theorem 6, we do not present a formal algorithm box here. Instead, we simply describe it as follows:

1. scan (from largest to smallest) through vector $\mathbf{q}$ to get cumulative summation $\sum_{j=a}^K \exp(\frac{\mathbf{q}_{[j]}}{\lambda} - 1)$ for all possible values of a.
2. apply the theorem 6 to calculate $[\mathbf{p}(a)]_{[a]}$ for all possible values of a. It is obvious that the cost is linear with K.

It is worth noting that the bottleneck for computing the analytical solution is sorting for $\mathbf{q}$, which usually costs O$(K \log K)$. Hence, the computation of the LDR-k-KL loss can be efficient conducted.

Proof. Apply Lagrangian multiplier:

$$\max_{\mathbf{p}\geq 0} \min_{\beta\geq 0, \boldsymbol{\gamma}\geq 0} \mathbf{p}^\top \frac{\mathbf{q}}{\lambda} - \sum_i \mathbf{p}_i \log K\mathbf{p}_i + \beta(1 - \sum_i \mathbf{p}_i) + \sum_i \gamma_i(\frac{1}{k} - \mathbf{p}_i)$$

Stationary condition of $\mathbf{p}$:

$$K\mathbf{p}_i^* = \exp(\frac{\mathbf{q}_i}{\lambda} - \beta - \gamma_i - 1)$$

Dual form:

$$\min_{\beta\geq 0, \boldsymbol{\gamma}\geq 0} \sum_i \frac{1}{K} \exp(\frac{\mathbf{q}_i}{\lambda} - \beta - \gamma_i - 1) + \beta + \sum_i \frac{\gamma_i}{k}$$

Stationary condition for β and $\boldsymbol{\gamma}$:

$$\gamma_i^* = \max\big(\log \frac{k}{K} \exp(\frac{\mathbf{q}_i}{\lambda} - \beta^* - 1), 0\big)$$

$$\beta^* = \max\big(\log \sum_i \frac{1}{K} \exp(\frac{\mathbf{q}_i}{\lambda} - \gamma_i^* - 1), 0\big)$$

Checkpoint I: for the $[a]$ that $\boldsymbol{\gamma}^*_{[a]} = 0$ and $\boldsymbol{\gamma}^*_{[a-1]} > 0$, where $2 \leq a \leq k+1$, (or $a = 1$, $\boldsymbol{\gamma}^*_{[a-1]}$ is undefined):

- We first consider the case when $\beta^* = 0$, then the $[\mathbf{p}(a)]^*_{[i]} = \frac{1}{k}$ for $i < a$, and $[\mathbf{p}(a)]^*_{[i]} = \frac{1}{K}\exp(\frac{\mathbf{q}_{[i]}}{\lambda} - 1)$ for $i \geq a$.
- Otherwise, if $\beta^* > 0$:

$$
\begin{aligned}
[\mathbf{p}(a)]^*_{[a]} &= \frac{1}{K}\exp(\frac{\mathbf{q}_{[a]}}{\lambda} - \beta^* - 1) = \frac{\exp(\frac{\mathbf{q}_{[a]}}{\lambda} - 1)}{\sum_j \exp(\frac{\mathbf{q}_j}{\lambda} - \gamma^*_j - 1)} \\
&= \frac{\exp(\frac{\mathbf{q}_{[a]}}{\lambda} - 1)}{\sum_{j=a}^K \exp(\frac{\mathbf{q}_{[j]}}{\lambda} - \gamma^*_{[j]} - 1)} \times \frac{\sum_{j=a}^K \exp(\frac{\mathbf{q}_{[j]}}{\lambda} - \gamma^*_{[j]} - 1)}{\sum_{j=1}^K \exp(\frac{\mathbf{q}_j}{\lambda} - \gamma^*_j - 1)} \\
&= \frac{\exp(\frac{\mathbf{q}_{[a]}}{\lambda} - 1)}{\sum_{j=a}^K \exp(\frac{\mathbf{q}_{[j]}}{\lambda} - 1)}(1 - \frac{a-1}{k})
\end{aligned}
$$

  It is similar to show for any i that $i > a$:

$$
[\mathbf{p}(a)]^*_{[i]} = \frac{\exp(\frac{\mathbf{q}_{[i]}}{\lambda} - 1)}{\sum_{j=a}^K \exp(\frac{\mathbf{q}_{[j]}}{\lambda} - 1)}(1 - \frac{a-1}{k})
$$

  For those $i < a$ that $\gamma^*_{[i]} > 0$:

$$
\begin{aligned}
[\mathbf{p}(a)]^*_{[i]} &= \frac{1}{K}\exp(\frac{\mathbf{q}_i}{\lambda} - \beta^* - \gamma^*_i - 1) \\
&= \frac{\exp(\frac{\mathbf{q}_i}{\lambda} - \gamma^*_i - 1)}{K\exp(\beta^*)} \\
&= \frac{\frac{K}{k}\exp(\beta^* - \frac{\mathbf{q}_i}{\lambda} + 1)\exp(\frac{\mathbf{q}_i}{\lambda} - 1)}{K\exp(\beta^*)} = \frac{1}{k}
\end{aligned}
$$

  So we can get the analytical solution as far as we know a subject to $\gamma^*_{[a]} = 0$ and $\gamma^*_{[a-1]} > 0$. The $[\mathbf{p}(a)]^*_{[i]} = \frac{1}{k}$ for $i < a$, and $[\mathbf{p}(a)]^*_{[i]} = \frac{\exp(\frac{\mathbf{q}_{[i]}}{\lambda} - 1)}{\sum_{j=a}^K \exp(\frac{\mathbf{q}_{[j]}}{\lambda} - 1)}(1 - \frac{a-1}{k})$ for $i \geq a$.

Checkpoint II: next, we show that as far as we find the smallest a such that $\boldsymbol{\gamma}^*_{[a]} = 0$ and $\boldsymbol{\gamma}^*_{[a-1]} > 0$, then it is the optimal solution.

- Suppose $a' < a$: because $\forall a' \;\; s.t. \;\; k+1 \geq a > a' \geq 1$, assume $[\mathbf{p}(a')]_{[a']} \leq \frac{1}{k} \implies a'$ is another smaller valid a, which violates the pre-condition; therefore, $[\mathbf{p}(a')]_{[a']} > [\mathbf{p}(a)]_{[a']} = \frac{1}{k}$, the a^* must lead to a $\mathbf{p}^*$ that violate the Ω^k constrain, hence can't be the optimal solution, contradiction.
- Suppose $a' > a$:

  i) if $\beta^*(a) = 0$, then by pre-condition $[\boldsymbol{\gamma}(a)]^*_{[a]} = 0$ and $[\boldsymbol{\gamma}(a')]^*_{[a]} > 0$, consequently $\beta^*(a') = 0$, which deviates from the optimality of the objective function:

$$
\min_{\beta \geq 0, \boldsymbol{\gamma} \geq 0} \sum_i \frac{1}{K}\exp(\frac{\mathbf{q}_i}{\lambda} - \beta - \boldsymbol{\gamma}_i - 1) + \beta + \sum_i \frac{\boldsymbol{\gamma}_i}{k}
$$

  (notice that $[\boldsymbol{\gamma}(a)]^*_{[i]} = [\boldsymbol{\gamma}(a')]^*_{[i]} > 0, \forall i < a$ in order to hold the $\Omega(k)$ constrain).

  ii) if $\beta^*(a) > 0$, then $[\boldsymbol{\gamma}(a)]^*_{[a]} = 0$ and $[\boldsymbol{\gamma}(a')]^*_{[a]} > 0 \implies \beta^*(a') < \beta^*(a) \implies \forall i \geq a', \; [\mathbf{p}(a')]_{[i]} > [\mathbf{p}(a)]_{[i]}$.

  On the other hand, $\forall j < a', \; [\mathbf{p}(a')]_{[j]} = \frac{1}{k} \geq [\mathbf{p}(a)]_{[j]} \implies \sum_{i=a'}^K [\mathbf{p}(a')]_{[i]} \leq 1 - \sum_{i=1}^{a'-1} [\mathbf{p}(a')]_{[i]} \leq 1 - \sum_{i=1}^{a'-1} [\mathbf{p}(a)]_{[i]} = \sum_{i=a'}^K [\mathbf{p}(a)]_{[i]}$, which is contradictory with $\forall i \geq a', \; [\mathbf{p}(a')]_{[i]} > [\mathbf{p}(a)]_{[i]}$.

□

J. Extensive Discussion for Loss Functions

From our theorem 5, we conclude that MAE, RCE, NCE and RLL are not All-k consistent. We have already proved the All-k consistency for LDR loss family by theorem 1 and theorem 2 (e.g. LDR-KL loss).

Next, we provide proofs for the All-k consistency of GCE ($0 \leq q < 1$), SCE ($0 \leq \alpha < 1$) and MSE.

Proof of All-k consistency for GCE ($0 \leq q < 1$)

To make the presentation clearer, we rename the q parameter in GCE loss function as λ, in order to not confuse with underlying class distribution notation $\mathbf{q}$. Objective:

$$\min_{\mathbf{p}\in\Delta_K} L_{\psi_{\text{GCE}}}(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{K} q_i \frac{1-p_i^{\lambda}}{\lambda}$$

Proof. We will first show $\mathbf{p}^*$ is rank consistent with $\mathbf{q}$; then, show $\mathbf{f}^*$ is rank consistent with $\mathbf{p}^*$; finally, it is easy to see that $\mathbf{f}^*$ is rank consistent with $\mathbf{q}$ by transitivity rule.

By Lagrangian multiplier:

$$\min_{\mathbf{p}} \max_{\alpha,\beta_i\geq 0} \sum_{i=1}^{K} q_i \frac{1-p_i^{\lambda}}{\lambda} + \alpha(\sum_{i=1}^{K} p_i - 1) - \sum_{i=1}^{K} \beta_i p_i$$

The stationary condition for $\mathbf{p}$:

$$p_i^* = \left(\frac{q_i}{\alpha-\beta_i}\right)^{\frac{1}{1-\lambda}}$$

Notice that the underlying class distribution $\mathbf{q} \in \Delta_K$, and $0 \leq \lambda < 1 \implies 1 \leq \frac{1}{1-\lambda}$; moreover, with primal feasibility for KKT condition, $p_i^* \geq 0$ and $\sum_{i=1}^{K} p_i^* = 1$ (it also practically holds because the model requires the output $\mathbf{p} \in \Delta_K$, e.g. softmax); then, consider the complementary slackness: $p_i^* > 0 \implies \beta_i^* = 0$; besides, $p_i^* \geq 0, q_i \geq 0 \implies \alpha > 0$; therefore, $p_i^* \propto q_i^{\frac{1}{1-\lambda}}$ and consequently $\mathbf{p}^*$ is rank consistent with $\mathbf{q}$.

Next, it is clear that $p_i = \frac{\exp(f_i)}{\sum_k \exp(f_k)} \implies \exp(f_i^*) \propto p_i^*$; therefore, $\mathbf{f}^*$ is rank consistent with $\mathbf{p}^*$ (notice that the proof is not restricted with softmax). By transitivity, $\mathbf{f}^*$ is rank consistent with $\mathbf{q}$. □

Proof of All-k consistency for SCE ($0 < \alpha \leq 1$)

Without loss of generality, we consider the case $A = -1$ but it could be generalized to any $A < 0$. We rename the α parameter for SCE loss as $1-\lambda$ for a better presentation. Objective:

$$\min_{\mathbf{p}\in\Delta_K} L_{\psi_{\text{SCE}}}(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{K} q_i \left(-(1-\lambda)\log p_i + \lambda(1-p_i)\right)$$

By Lagrangian multiplier:

$$\min_{\mathbf{p}} \max_{\alpha,\beta_i\geq 0} \sum_{i=1}^{K} q_i \left(-(1-\lambda)\log p_i + \lambda(1-p_i)\right) + \alpha(\sum_{i=1}^{K} p_i - 1) - \sum_{i=1}^{K} \beta_i p_i$$

The stationary condition for $\mathbf{p}$:

$$p_i^* = \frac{(1-\lambda)q_i}{\alpha-\beta_i-\lambda q_i}$$

$$\frac{1}{p_i^*} = \frac{\alpha-\beta_i-\lambda q_i}{(1-\lambda)q_i} = \frac{\alpha-\beta_i}{(1-\lambda)q_i} - \frac{\lambda}{1-\lambda}, \quad 0 < p_i^* < 1$$

By primal feasibility for KKT condition, $p_i^* \geq 0$ and $\sum_{i=1}^{K} p_i^* = 1$. By complementary slackness: $p_i^* > 0 \implies \beta_i^* = 0$. Besides, $p_i^* \geq 0, q_i \geq 0 \implies \alpha > 0$. Notice that $0 < 1-\lambda \leq 1$. Therefore, $(\frac{1}{p_i^*} + \frac{\lambda}{1-\lambda}) \propto \frac{1}{q_i}$, and consequently $\mathbf{p}^*$ is rank consistent with $\mathbf{q}$.

Next, it is obvious that $p_i = \frac{\exp(f_i)}{\sum_k \exp(f_k)} \implies \exp(f_i^*) \propto p_i^*$; therefore, $\mathbf{f}^*$ is rank consistent with $\mathbf{p}^*$. By transitivity, $\mathbf{f}^*$ is rank consistent with $\mathbf{q}$.

Notice that the proof is for the empirical form of SCE. For the theoretical form: $L_{\psi_{\text{SCE}}}(\mathbf{p},\mathbf{q}) = H(\mathbf{q},\mathbf{p}) + H(\mathbf{p},\mathbf{q})$, where $H(\mathbf{q},\mathbf{p}) = -\mathrm{E}_q[\log p]$ denotes cross entropy, it can be proved with the similar logic.

Proof of All-k consistency for MSE

Objective:

$$\min_{\mathbf{p}\in\Delta_K} L_{\psi_{\text{MSE}}}(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{K} q_i(1-2p_i+\|\mathbf{p}\|_2^2)$$

By Lagrangian multiplier:

$$\min_{\mathbf{p}}\max_{\alpha,\beta_i\geq 0} \sum_{i=1}^{K} q_i(1-2p_i+\|\mathbf{p}\|_2^2) + \alpha(\sum_{i=1}^{K} p_i - 1) - \sum_{i=1}^{K}\beta_i p_i$$

The stationary condition for $\mathbf{p}$:

$$p_i^* = q_i + \frac{\beta_i - \alpha}{2}$$

By the primal feasibility for KKT condition, we have $p_i^* \geq 0$ and $\sum_{i=1}^K p_i^* = 1$. By complementary slackness: $p_i^* > 0 \implies \beta_i^* = 0$. Furthermore, $p_i^* = 0$, we can prove $\beta_i^* = 0$ holds; otherwise, $\exists j,\ s.t.\ p_j^* = 0, \beta_j^* > 0$, with primal feasibility $\sum_{i=1}^K \left(q_i + \frac{\beta_i^* - \alpha^*}{2}\right) = 1 \implies \alpha^* > 0$ given that $\sum_{i=1}^K q_i = 1$. As a consequence, for the $p_i^* = 0$, we conclude $p_i^* \leq q_i$; for the $p_i^* > 0$, because $\beta_i^* = 0, \alpha^* > 0$, we conclude $p_i^* = q_i + \frac{\beta_i^* - \alpha^*}{2} < q_i$; therefore, $\sum_{i=1}^K p_i^* < \sum_{i=1}^K q_i = 1$, a contraction. Hence, $\mathbf{p}^* = \mathbf{q}$ and is rank consistent with $\mathbf{q}$.

Next, it is obvious that $p_i = \frac{\exp(f_i)}{\sum_k \exp(f_k)} \implies \exp(f_i^*) \propto p_i^*$; therefore, $\mathbf{f}^*$ is rank consistent with $\mathbf{p}^*$. By transitivity, $\mathbf{f}^*$ is rank consistent with $\mathbf{q}$.

K. Dataset Statistics

Table 5. statistics for the benchmark datasets

Dataset	# of samples	# of features	# of classes
ALOI	108000	128	1,000
News20	15935	62060	20
Letter	15000	16	26
Vowel	990	10	11
Kuzushiji-49	232365 / 38547	28x28	49
CIFAR-100	50000 / 10000	32x32x3	100
Tiny-ImageNet	100000 / 10000	64x64x3	200

L. More Experimental Results

L.1. Relationship between λ and KL divergence values $\text{KL}(\mathbf{p}, 1/K)$ for ALDR-KL on Synthetic Data

We present the averaged learned λ_t and $\text{KL}(\mathbf{p}_t, \frac{\mathbf{1}}{K})$ values for ALDR-KL loss at Figure 3. It justifies our motivation that we prefer to give the uncertain data with larger λ value.

L.2. Comprehensive Accuracy Results on Benchmark datasets

The top-$1, 2, 3, 4, 5$ accuracy results are summarized at Table 6, 7, 8, 9, 10, 11, 12, 13, 14, 15. From the results, we can see that (i) the performance of ALDR-KL and LDR-KL are stable; (ii) MAE has very poor performance, which is consistent with many earlier works; (ii) CE is still competitive in most cases.

L.3. Setups for Mini-Webvision Experiments

We adopt the noisy Google resized version from Webvision version 1.0 as training dataset (only for the first 50 classes) (Li et al., 2017) and do evaluation on the ILSVRC 2012 validation data (only for the first 50 classes). The total training epochs is set as 250. The weight decay is set as 3e-5. SGD optimizer is utilized with nesterov momentum as 0.9, initial learning rate as 0.4 that is epoch-wise decreased by 0.97. Typical data augmentations including color jittering, random width/height shift and random horizontal flip are applied. For our proposed LDR-KL and ALDR-KL, we apply SoftPlus function to make model output $\mathbf{f}$ to be positive. The results are presented at Table 3.

Table 6. Testing Top-1 Accuracy (%) with mean and standard deviation for tabular dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Vowel	ALDR-KL	71.3(3.1)	**58.2(6.38)**	39.8(6.25)	**15.2(6.13)**	69.1(3.36)	**61.2(2.08)**	39.4(3.06)
	LDR-KL	**74.1(2.18)**	56.4(5.32)	35.4(4.65)	13.7(4.02)	**73.1(2.68)**	58.6(6.39)	**41.0(3.7)**
	CE	63.0(3.48)	53.9(3.42)	**41.0(2.9)**	14.3(4.88)	65.9(3.52)	56.4(4.88)	35.2(2.34)
	SCE	64.4(3.69)	55.4(4.35)	37.4(4.19)	11.9(2.96)	65.5(2.16)	55.6(3.94)	34.7(3.17)
	GCE	64.2(4.41)	52.1(4.81)	39.2(2.51)	11.9(6.65)	64.6(2.47)	56.4(3.58)	39.0(3.23)
	TGCE	64.6(3.61)	52.1(4.81)	38.6(2.74)	11.9(6.65)	64.6(2.47)	56.4(3.58)	38.8(3.48)
	WW	58.0(4.22)	47.9(1.51)	34.1(6.83)	11.3(6.56)	57.0(1.87)	50.9(2.08)	37.2(3.4)
	JS	62.6(2.3)	57.4(2.42)	39.0(3.1)	14.1(1.92)	63.8(4.06)	53.5(3.5)	34.5(2.51)
	CS	70.1(3.23)	49.3(2.74)	32.9(6.41)	9.7(3.23)	65.5(2.74)	56.2(3.97)	36.8(3.42)
	RLL	61.8(2.25)	56.4(2.06)	38.6(3.75)	9.29(5.4)	63.6(0.903)	49.1(2.27)	34.9(3.87)
	NCE+RCE	61.0(7.58)	47.3(5.21)	32.7(3.04)	13.3(6.77)	56.2(4.89)	45.5(3.67)	32.5(7.01)
	NCE+AUL	54.1(3.92)	53.7(8.74)	29.9(1.87)	12.3(5.09)	50.1(6.21)	46.7(3.75)	32.1(6.62)
	NCE+AGCE	57.6(3.78)	51.3(10.3)	32.7(3.92)	12.1(4.95)	57.6(5.46)	45.0(3.17)	31.7(3.92)
	MSE	56.0(1.37)	43.2(3.58)	29.3(1.69)	9.9(4.4)	52.5(2.3)	45.0(5.01)	32.7(1.03)
	MAE	10.1(2.12)	10.1(2.12)	9.09(2.71)	8.48(2.36)	8.69(4.27)	8.69(4.27)	9.09(4.04)
Letter	ALDR-KL	79.9(0.254)	**76.1(0.445)**	68.6(0.543)	**36.1(2.12)**	**79.2(0.504)**	**77.2(0.463)**	50.8(0.667)
	LDR-KL	**80.1(0.124)**	75.7(0.847)	**68.8(0.676)**	36.0(2.19)	78.8(0.178)	75.8(0.434)	**51.6(0.893)**
	CE	74.2(0.538)	68.2(0.414)	56.9(0.495)	16.1(1.81)	73.6(0.183)	70.5(0.767)	48.1(1.29)
	SCE	74.0(0.585)	67.9(0.431)	55.9(0.546)	15.0(1.66)	73.2(0.341)	70.5(0.969)	48.1(1.76)
	GCE	73.2(0.428)	67.0(0.409)	54.2(0.494)	13.7(1.96)	72.7(0.392)	69.8(0.877)	47.1(1.87)
	TGCE	73.2(0.44)	66.3(0.284)	54.3(1.18)	14.1(2.1)	72.7(0.392)	69.8(0.888)	46.4(1.45)
	WW	65.3(1.04)	61.6(0.969)	56.1(1.26)	14.9(1.6)	65.2(0.222)	62.2(0.912)	42.1(1.66)
	JS	72.6(0.361)	65.9(0.518)	45.0(1.55)	8.33(2.37)	71.3(0.325)	68.5(0.741)	46.6(1.96)
	CS	71.5(0.284)	65.0(0.739)	60.6(0.606)	24.4(3.58)	67.7(1.04)	61.5(1.05)	43.0(2.04)
	RLL	69.2(0.515)	57.8(0.501)	28.0(0.479)	5.1(1.35)	67.7(0.651)	61.8(0.598)	42.5(1.22)
	NCE+RCE	65.8(1.66)	60.2(2.46)	51.6(4.7)	19.6(3.2)	63.4(1.52)	49.5(1.94)	44.6(2.0)
	NCE+AUL	64.7(2.78)	58.3(1.84)	49.3(2.75)	13.1(1.31)	58.7(1.47)	49.2(2.71)	42.0(1.5)
	NCE+AGCE	57.5(1.0)	51.0(3.19)	40.1(4.16)	5.07(1.42)	54.0(2.57)	43.2(2.29)	38.6(2.82)
	MSE	27.1(2.17)	15.0(2.44)	11.7(1.99)	5.7(1.68)	20.9(0.429)	16.9(1.5)	15.6(1.76)
	MAE	5.03(1.22)	5.14(1.08)	4.86(1.27)	4.4(0.499)	5.04(1.2)	5.09(1.19)	5.05(1.23)
News20	ALDR-KL	91.6(0.161)	**88.2(0.376)**	77.9(1.0)	**22.4(2.39)**	**90.1(0.157)**	86.0(0.281)	63.9(0.882)
	LDR-KL	**92.3(0.256)**	**88.2(0.371)**	**78.1(1.29)**	18.0(1.67)	89.7(0.472)	**86.2(0.291)**	**64.3(1.12)**
	CE	62.0(1.0)	30.0(0.821)	14.5(1.64)	5.59(0.562)	53.4(0.936)	53.0(1.42)	49.7(0.671)
	SCE	57.3(1.11)	27.2(0.714)	14.0(1.6)	5.46(1.2)	51.5(0.651)	50.7(1.51)	47.8(0.931)
	GCE	46.0(0.526)	19.8(0.95)	12.8(1.34)	5.41(1.08)	46.2(0.847)	44.5(0.839)	41.4(1.05)
	TGCE	45.8(0.635)	20.8(1.38)	12.7(1.48)	5.41(1.08)	45.9(1.02)	44.1(0.973)	40.7(1.34)
	WW	53.2(3.31)	36.9(5.36)	23.7(2.16)	6.72(2.8)	51.6(1.81)	47.6(1.23)	42.8(0.875)
	JS	25.5(0.196)	12.2(0.714)	10.3(1.57)	4.7(0.675)	23.5(0.923)	19.5(1.23)	17.2(1.35)
	CS	77.5(2.51)	73.5(0.906)	55.3(1.65)	9.7(2.07)	75.4(1.85)	64.7(1.2)	54.4(0.713)
	RLL	11.4(1.41)	10.1(1.67)	7.39(1.28)	5.16(0.871)	10.9(1.01)	10.3(0.748)	10.2(0.754)
	NCE+RCE	90.4(0.213)	72.0(1.18)	28.8(2.17)	6.0(0.785)	86.4(2.56)	65.2(1.01)	63.0(0.945)
	NCE+AUL	63.1(2.0)	31.0(1.92)	15.0(0.265)	5.72(0.619)	54.1(1.87)	47.2(3.4)	44.0(2.32)
	NCE+AGCE	39.9(3.09)	19.3(1.78)	10.3(0.519)	5.78(0.738)	34.2(1.9)	29.8(3.19)	29.7(0.31)
	MSE	7.86(1.19)	8.35(1.43)	6.31(0.97)	4.99(0.484)	6.58(1.08)	8.39(3.02)	8.0(1.18)
	MAE	5.01(0.241)	5.24(0.155)	5.12(0.119)	4.88(0.703)	5.17(0.261)	4.99(0.336)	5.04(0.39)
ALOI	ALDR-KL	**94.4(0.103)**	89.2(0.054)	82.7(0.197)	32.7(0.307)	**91.9(0.163)**	79.4(0.206)	**45.7(0.273)**
	LDR-KL	92.7(0.101)	**89.3(0.105)**	**83.3(0.117)**	**32.9(0.544)**	91.1(0.13)	**83.2(0.221)**	44.5(0.497)
	CE	75.7(0.148)	65.8(0.298)	40.9(0.611)	3.93(0.254)	73.3(0.394)	61.5(0.259)	26.4(0.31)
	SCE	74.7(0.181)	63.8(0.312)	38.8(0.614)	3.74(0.307)	71.8(0.317)	59.6(0.241)	25.3(0.127)
	GCE	71.4(0.205)	56.5(0.151)	29.6(0.655)	2.64(0.324)	67.7(0.31)	52.7(0.492)	20.9(0.305)
	TGCE	70.2(0.389)	54.9(0.331)	28.5(0.78)	2.64(0.324)	66.2(0.068)	50.6(0.512)	20.0(0.154)
	WW	57.3(0.464)	43.5(0.349)	24.2(0.435)	2.4(0.315)	44.8(0.476)	23.5(0.179)	9.34(0.194)
	JS	3.53(0.595)	2.88(0.443)	1.72(0.338)	0.189(0.058)	3.45(0.488)	2.43(0.266)	1.63(0.282)
	CS	84.5(0.18)	73.4(0.113)	30.2(0.411)	1.63(0.23)	81.5(0.224)	75.6(0.215)	38.7(0.479)
	RLL	0.572(0.138)	0.413(0.145)	0.22(0.073)	0.15(0.061)	0.55(0.149)	0.381(0.136)	0.207(0.109)
	NCE+RCE	3.9(0.49)	2.73(0.155)	2.02(0.535)	0.269(0.109)	2.93(0.266)	1.87(0.209)	1.56(0.196)
	NCE+AUL	2.64(0.537)	1.81(0.278)	0.806(0.203)	0.157(0.055)	2.51(0.543)	1.84(0.248)	1.22(0.203)
	NCE+AGCE	2.26(0.429)	1.22(0.313)	0.432(0.105)	0.128(0.042)	2.03(0.494)	1.36(0.206)	0.826(0.193)
	MSE	0.217(0.084)	0.185(0.089)	0.172(0.059)	0.096(0.051)	0.2(0.09)	0.187(0.067)	0.137(0.102)
	MAE	0.152(0.067)	0.144(0.062)	0.124(0.047)	0.089(0.038)	0.152(0.067)	0.137(0.077)	0.139(0.051)

Table 7. Testing Top-1 Accuracy (%) with mean and standard deviation for image dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Kuzushiji-49	ALDR-KL	96.6(0.132)	94.5(0.248)	**91.5(0.154)**	**48.2(1.54)**	96.0(0.256)	**95.5(0.11)**	**54.4(5.84)**
	LDR-KL	96.6(0.06)	**94.8(0.145)**	91.2(0.295)	48.1(1.27)	95.9(0.089)	95.2(0.132)	54.1(3.69)
	CE	96.9(0.09)	94.0(0.102)	85.5(0.219)	29.4(1.45)	96.2(0.064)	95.2(0.098)	47.7(3.17)
	SCE	96.9(0.074)	94.4(0.186)	85.5(0.235)	29.1(0.836)	96.2(0.096)	95.2(0.17)	49.7(2.1)
	GCE	96.9(0.069)	94.5(0.161)	85.5(0.347)	28.6(0.684)	96.2(0.076)	95.2(0.085)	48.8(1.69)
	TGCE	96.8(0.019)	94.4(0.115)	85.5(0.221)	28.7(0.359)	96.2(0.05)	95.2(0.114)	48.2(2.22)
	WW	**97.1(0.091)**	90.0(0.341)	80.1(0.236)	26.3(1.45)	95.5(0.452)	94.5(0.381)	49.8(1.22)
	JS	96.8(0.052)	92.2(0.598)	87.4(0.742)	31.2(1.09)	94.8(0.723)	93.0(0.188)	49.3(2.71)
	CS	97.0(0.03)	86.8(3.0)	6.93(2.46)	2.1(0.087)	**96.3(0.2)**	91.2(0.729)	37.1(2.93)
	RLL	91.4(0.031)	90.5(0.082)	81.1(1.32)	30.6(0.866)	90.8(0.069)	86.6(1.05)	47.8(3.13)
	NCE+RCE	90.7(0.104)	90.0(0.134)	69.2(0.778)	8.48(1.08)	87.1(0.062)	81.2(0.877)	43.2(0.64)
	NCE+AUL	91.0(0.06)	85.3(0.923)	65.6(1.43)	6.9(1.16)	89.0(0.058)	83.0(0.905)	45.0(1.8)
	NCE+AGCE	90.8(0.088)	77.4(0.83)	58.2(2.15)	6.51(0.769)	88.1(0.873)	79.0(0.9)	42.6(1.33)
	MSE	82.6(1.65)	57.7(1.53)	35.8(1.56)	4.18(0.796)	63.6(1.7)	46.9(1.59)	28.4(1.14)
	MAE	2.09(0.062)	35.2(2.1)	20.0(3.03)	3.81(0.702)	39.1(2.01)	24.2(1.81)	18.2(1.33)
CIFAR100	ALDR-KL	**60.6(0.725)**	**51.9(0.378)**	36.0(0.495)	3.52(0.536)	57.5(0.584)	51.7(0.711)	30.0(1.23)
	LDR-KL	60.5(0.535)	51.8(0.75)	**36.3(0.698)**	3.7(0.544)	**57.6(0.574)**	51.2(0.397)	29.5(0.982)
	CE	59.5(0.545)	48.7(0.617)	32.0(0.941)	3.26(0.357)	57.0(0.61)	50.1(0.399)	29.5(0.7)
	SCE	59.5(0.773)	49.9(0.846)	30.7(1.03)	3.87(0.251)	57.0(0.245)	50.7(0.754)	30.0(0.488)
	GCE	60.1(0.558)	49.2(0.925)	31.9(0.831)	3.69(0.392)	57.0(0.92)	50.4(0.806)	**30.2(0.616)**
	TGCE	59.5(0.689)	49.8(0.555)	28.8(0.568)	3.59(0.398)	56.9(0.602)	**53.5(0.445)**	30.1(0.77)
	WW	50.2(0.625)	36.1(0.976)	21.3(0.71)	4.1(0.346)	48.2(0.505)	44.6(0.252)	25.9(0.505)
	JS	58.7(0.424)	52.2(0.794)	30.3(1.18)	3.5(0.438)	57.2(0.363)	49.5(1.06)	29.1(0.429)
	CS	11.0(0.404)	4.85(0.345)	1.59(0.217)	1.0(0.004)	8.07(0.333)	5.42(0.468)	3.71(0.194)
	RLL	42.5(1.95)	35.4(0.438)	18.3(1.16)	3.57(0.428)	40.3(0.855)	34.7(1.84)	17.9(1.21)
	NCE+RCE	13.5(1.22)	8.29(0.949)	5.03(0.563)	2.06(0.697)	12.5(1.28)	8.32(0.439)	4.6(0.429)
	NCE+AUL	11.5(0.368)	7.73(0.403)	5.79(0.125)	2.44(0.839)	9.9(0.792)	7.6(0.316)	4.33(0.189)
	NCE+AGCE	10.3(0.498)	7.52(0.298)	5.61(0.123)	2.03(0.896)	9.16(0.516)	7.51(0.277)	4.06(0.224)
	MSE	19.0(0.663)	10.1(0.545)	8.22(0.476)	**4.61(0.524)**	16.0(0.579)	9.83(0.358)	5.99(1.02)
	MAE	1.07(0.152)	1.05(0.063)	0.918(0.153)	1.01(0.182)	1.03(0.093)	1.09(0.127)	0.968(0.28)
Tiny-ImageNet	ALDR-KL	**47.8(0.226)**	**38.4(0.136)**	**23.7(0.216)**	1.3(0.145)	44.9(0.167)	38.1(0.634)	21.6(0.39)
	LDR-KL	47.6(0.39)	38.4(0.521)	23.4(0.563)	1.29(0.201)	45.1(0.171)	37.4(0.112)	**22.1(0.333)**
	CE	47.7(0.375)	36.3(0.246)	14.5(0.316)	1.23(0.187)	45.5(0.149)	41.5(0.262)	22.0(0.544)
	SCE	47.9(0.286)	36.1(0.175)	14.0(0.453)	**1.53(0.449)**	45.4(0.254)	41.5(0.244)	22.1(0.345)
	GCE	47.8(0.397)	36.1(0.327)	11.9(0.572)	0.968(0.281)	**45.7(0.099)**	**41.6(0.296)**	22.0(0.267)
	TGCE	45.7(0.106)	31.7(0.367)	10.7(0.564)	1.06(0.271)	43.3(0.324)	39.8(0.227)	19.7(0.362)
	WW	31.3(0.159)	20.7(0.354)	10.2(0.418)	0.99(0.109)	28.4(0.204)	24.9(0.631)	13.9(0.283)
	JS	39.9(0.405)	21.6(0.744)	4.26(0.235)	1.11(0.294)	35.5(0.334)	24.1(0.21)	10.1(0.786)
	CS	0.968(0.123)	0.672(0.127)	0.496(0.008)	0.502(0.004)	0.81(0.078)	0.66(0.07)	0.514(0.02)
	RLL	11.1(1.14)	3.9(0.566)	2.42(0.25)	0.648(0.242)	7.47(0.53)	3.53(0.18)	2.27(0.301)
	NCE+RCE	1.71(0.214)	1.5(0.119)	1.21(0.041)	0.802(0.107)	1.65(0.131)	1.55(0.19)	1.05(0.2)
	NCE+AUL	2.05(0.117)	1.66(0.094)	1.44(0.121)	0.866(0.234)	1.9(0.206)	1.66(0.226)	1.21(0.394)
	NCE+AGCE	1.98(0.145)	1.74(0.191)	1.51(0.166)	0.914(0.316)	1.97(0.221)	1.67(0.098)	1.18(0.235)
	MSE	2.53(0.179)	1.29(0.104)	0.998(0.087)	0.72(0.125)	1.29(0.071)	1.24(0.137)	1.1(0.163)
	MAE	0.548(0.083)	0.996(0.138)	0.838(0.087)	0.732(0.136)	1.07(0.165)	1.07(0.175)	1.01(0.189)

Table 8. Testing Top-2 Accuracy (%) with mean and standard deviation for tabular dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Vowel	ALDR-KL	**93.9(2.3)**	**83.0(2.67)**	**61.0(4.27)**	23.4(5.97)	**92.7(1.96)**	**93.3(2.27)**	**86.3(1.98)**
	LDR-KL	93.5(1.76)	81.4(3.29)	60.6(4.52)	23.6(5.44)	91.9(1.11)	90.1(1.74)	86.1(1.62)
	CE	88.7(1.49)	77.8(2.12)	57.8(1.34)	19.8(5.37)	90.3(2.83)	85.9(3.5)	82.6(3.52)
	SCE	88.5(1.51)	77.4(2.27)	57.2(6.6)	21.0(3.52)	88.5(4.07)	86.1(3.09)	80.2(2.08)
	GCE	88.7(1.18)	79.0(1.49)	58.0(2.6)	20.4(6.17)	88.9(4.74)	86.3(3.42)	80.2(2.75)
	TGCE	88.5(1.03)	79.0(1.49)	57.6(2.3)	20.4(6.17)	88.3(4.5)	86.3(3.42)	79.8(3.06)
	WW	84.6(2.06)	75.8(3.78)	55.2(4.36)	**25.5(3.91)**	86.7(3.09)	82.8(3.44)	76.0(3.34)
	JS	87.1(2.16)	81.2(1.51)	59.4(2.06)	19.8(7.02)	84.2(3.81)	85.1(2.06)	78.6(2.67)
	CS	88.7(1.85)	74.3(3.59)	49.1(8.75)	20.8(5.29)	84.6(1.85)	82.0(3.8)	70.9(2.34)
	RLL	82.2(1.76)	77.6(2.74)	57.4(3.58)	22.4(6.27)	83.2(2.36)	83.0(1.96)	77.0(3.28)
	NCE+RCE	73.7(3.61)	70.1(1.51)	47.5(3.0)	20.6(6.84)	73.5(1.85)	71.5(3.69)	74.5(0.756)
	NCE+AUL	73.9(2.25)	67.3(3.92)	43.0(6.57)	22.2(5.15)	72.9(0.756)	74.3(1.64)	71.9(2.25)
	NCE+AGCE	74.1(2.44)	65.3(3.1)	42.4(5.79)	18.0(6.04)	72.5(0.756)	72.9(0.99)	70.7(1.28)
	MSE	72.7(2.47)	67.5(2.34)	44.4(4.14)	23.4(4.35)	75.4(1.37)	69.3(2.36)	68.3(4.07)
	MAE	18.4(3.16)	18.4(3.16)	22.2(3.44)	22.4(3.4)	18.4(3.16)	19.4(2.74)	19.0(1.74)
Letter	ALDR-KL	**86.3(0.366)**	**84.7(0.318)**	79.8(0.733)	50.6(1.5)	**86.4(0.588)**	**84.9(0.231)**	**81.4(0.56)**
	LDR-KL	86.1(0.16)	84.3(0.416)	**80.2(0.519)**	**50.9(1.78)**	85.9(0.073)	84.0(0.446)	80.3(0.543)
	CE	83.5(0.208)	80.5(0.308)	72.5(1.0)	24.6(3.34)	83.3(0.271)	81.3(0.222)	75.6(0.331)
	SCE	83.4(0.139)	80.2(0.336)	71.8(1.08)	25.1(3.52)	83.4(0.224)	81.2(0.29)	75.3(0.452)
	GCE	82.9(0.275)	80.0(0.22)	70.5(0.924)	22.7(5.18)	82.7(0.132)	80.8(0.212)	74.8(0.35)
	TGCE	82.9(0.166)	79.9(0.248)	70.1(1.41)	22.7(5.18)	82.7(0.13)	80.6(0.232)	74.7(0.349)
	WW	81.2(0.314)	79.9(0.3)	75.1(0.92)	26.0(1.8)	80.9(0.236)	77.1(0.213)	70.5(0.408)
	JS	82.0(0.277)	78.7(0.548)	62.7(0.947)	16.1(2.3)	81.8(0.33)	79.5(0.354)	72.8(0.376)
	CS	82.2(0.256)	78.4(0.467)	73.9(1.06)	36.9(5.1)	81.0(0.384)	77.9(0.585)	71.3(0.576)
	RLL	79.7(0.635)	70.9(0.946)	39.0(1.0)	9.53(1.78)	78.9(0.379)	75.0(0.61)	66.3(0.624)
	NCE+RCE	69.0(1.46)	63.0(2.12)	53.9(4.76)	20.8(3.81)	65.9(1.98)	51.7(1.79)	47.2(1.54)
	NCE+AUL	67.6(3.46)	61.1(1.91)	51.8(2.97)	12.4(3.06)	61.2(1.64)	51.8(3.07)	44.4(1.32)
	NCE+AGCE	60.0(1.3)	53.7(3.3)	41.6(4.46)	10.6(1.22)	56.2(2.43)	45.0(2.56)	40.7(2.99)
	MSE	31.9(0.666)	24.5(3.14)	18.9(2.76)	9.33(2.23)	27.9(3.09)	26.7(2.26)	26.1(2.76)
	MAE	8.98(0.98)	8.87(0.827)	8.86(0.854)	7.86(0.857)	9.09(1.02)	9.05(1.07)	9.06(1.08)
News20	ALDR-KL	96.4(0.294)	**94.3(0.176)**	86.4(0.635)	30.3(4.97)	95.2(0.214)	93.9(0.474)	92.5(0.494)
	LDR-KL	**96.6(0.2)**	94.2(0.243)	**86.6(0.929)**	**31.3(3.8)**	**95.9(0.229)**	**94.2(0.328)**	**93.0(0.4)**
	CE	78.9(0.557)	46.7(2.5)	22.2(0.871)	9.73(1.16)	76.4(0.627)	77.6(0.825)	76.3(0.699)
	SCE	75.7(0.425)	41.5(1.61)	20.9(1.52)	10.2(0.544)	73.8(0.705)	75.1(1.2)	73.7(0.974)
	GCE	67.7(1.02)	34.1(1.32)	18.7(1.03)	9.89(0.675)	66.6(1.08)	67.8(1.29)	66.6(1.41)
	TGCE	66.8(1.08)	34.1(1.36)	18.7(1.29)	9.89(0.675)	66.1(1.57)	66.7(1.52)	65.3(1.56)
	WW	71.3(1.49)	49.2(4.92)	32.4(2.38)	13.2(3.9)	69.5(1.48)	71.0(1.72)	69.3(1.18)
	JS	35.1(0.868)	19.3(0.16)	14.5(1.25)	11.1(1.15)	34.2(1.54)	31.4(1.02)	31.1(1.47)
	CS	87.3(1.6)	87.6(0.527)	69.2(1.1)	13.9(1.28)	86.6(1.62)	81.6(1.45)	78.8(1.7)
	RLL	18.3(0.375)	15.2(1.72)	15.3(0.738)	9.85(0.707)	18.8(1.47)	18.6(0.291)	18.2(0.759)
	NCE+RCE	94.5(0.263)	78.7(2.01)	32.5(2.11)	10.4(0.362)	92.4(2.44)	70.9(1.38)	66.7(0.689)
	NCE+AUL	65.0(2.38)	31.8(1.72)	19.0(1.52)	11.4(1.0)	55.7(1.96)	49.1(3.02)	46.7(1.9)
	NCE+AGCE	40.8(3.28)	20.2(1.26)	14.9(1.39)	11.8(0.839)	35.8(0.718)	30.8(3.36)	30.9(0.703)
	MSE	16.3(2.1)	13.4(0.99)	12.3(1.45)	9.37(0.252)	14.5(1.1)	15.1(1.68)	13.4(0.983)
	MAE	10.1(0.37)	10.1(0.476)	10.4(0.718)	10.2(0.641)	10.2(0.333)	10.2(0.459)	10.0(0.361)
ALOI	ALDR-KL	**97.5(0.065)**	**93.8(0.119)**	88.5(0.191)	40.4(0.881)	**96.0(0.044)**	**92.5(0.098)**	**87.6(0.179)**
	LDR-KL	96.3(0.126)	93.7(0.046)	**89.0(0.099)**	**40.6(0.671)**	95.0(0.063)	91.5(0.184)	85.6(0.185)
	CE	83.4(0.161)	76.5(0.178)	53.2(0.725)	6.14(0.299)	81.8(0.201)	72.9(0.215)	44.0(0.439)
	SCE	82.7(0.168)	75.0(0.2)	51.3(0.728)	6.37(0.36)	80.9(0.153)	71.8(0.313)	42.0(0.235)
	GCE	80.2(0.225)	69.1(0.329)	41.1(0.469)	4.63(0.463)	77.7(0.189)	65.2(0.416)	33.8(0.419)
	TGCE	79.1(0.288)	67.7(0.345)	39.9(0.68)	4.63(0.463)	76.6(0.183)	63.7(0.277)	32.3(0.439)
	WW	71.3(0.462)	58.4(0.513)	37.3(0.474)	4.79(0.159)	58.8(0.507)	36.2(0.263)	16.6(0.255)
	JS	5.75(0.464)	5.06(0.549)	2.84(0.399)	0.267(0.079)	5.47(0.58)	4.24(0.419)	2.74(0.351)
	CS	89.4(0.151)	80.6(0.246)	37.8(0.524)	2.76(0.036)	87.4(0.07)	83.0(0.081)	63.3(0.467)
	RLL	1.15(0.278)	0.756(0.144)	0.404(0.14)	0.28(0.071)	0.993(0.203)	0.763(0.149)	0.578(0.099)
	NCE+RCE	4.47(0.236)	4.36(0.224)	3.32(0.498)	0.433(0.178)	4.6(0.105)	3.44(0.237)	2.4(0.303)
	NCE+AUL	4.58(0.365)	3.01(0.42)	1.38(0.339)	0.222(0.058)	4.13(0.504)	2.99(0.308)	2.04(0.355)
	NCE+AGCE	3.62(0.488)	1.98(0.42)	0.887(0.179)	0.23(0.04)	3.12(0.648)	2.18(0.348)	1.47(0.274)
	MSE	0.483(0.117)	0.38(0.076)	0.343(0.071)	0.222(0.045)	0.441(0.123)	0.372(0.08)	0.38(0.055)
	MAE	0.33(0.087)	0.313(0.087)	0.302(0.08)	0.228(0.039)	0.313(0.087)	0.289(0.093)	0.306(0.097)

Table 9. Testing Top-2 Accuracy (%) with mean and standard deviation for image dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Kuzushiji-49	ALDR-KL	98.2(0.072)	**97.1(0.122)**	94.1(1.39)	**56.2(1.17)**	97.2(0.068)	96.5(0.306)	95.8(0.159)
	LDR-KL	98.3(0.022)	97.0(0.144)	**95.0(0.136)**	55.8(1.09)	97.2(0.132)	96.6(0.079)	95.9(0.207)
	CE	98.4(0.105)	96.9(0.091)	89.3(0.318)	41.5(2.13)	97.4(0.055)	96.7(0.136)	96.1(0.112)
	SCE	98.4(0.133)	97.1(0.127)	89.1(0.335)	41.0(1.29)	97.4(0.034)	96.7(0.069)	96.1(0.1)
	GCE	98.5(0.089)	97.2(0.104)	89.1(0.315)	38.5(1.23)	97.4(0.057)	96.7(0.062)	96.1(0.09)
	TGCE	98.4(0.047)	97.0(0.261)	89.2(0.243)	39.9(1.58)	97.4(0.071)	96.7(0.107)	96.1(0.081)
	WW	**98.6(0.129)**	96.1(0.127)	87.7(0.35)	38.5(1.47)	**97.6(0.055)**	**97.1(0.067)**	**96.8(0.079)**
	JS	98.3(0.02)	94.5(0.694)	90.2(0.839)	39.8(0.984)	96.4(0.902)	94.4(0.11)	93.8(0.101)
	CS	98.3(0.033)	90.5(2.17)	9.45(2.2)	4.16(0.157)	97.5(0.148)	94.1(0.706)	60.0(2.78)
	RLL	92.7(0.063)	92.2(0.068)	83.4(1.25)	37.5(0.809)	92.3(0.046)	89.6(0.896)	90.4(0.946)
	NCE+RCE	92.1(0.106)	91.7(0.157)	71.1(0.649)	12.0(1.47)	89.0(0.282)	84.3(0.436)	72.3(1.63)
	NCE+AUL	92.3(0.087)	86.8(0.868)	67.4(1.33)	12.9(1.0)	90.3(0.041)	87.0(0.349)	84.3(0.397)
	NCE+AGCE	92.1(0.039)	78.9(0.777)	59.1(1.52)	12.2(0.945)	89.4(0.867)	84.3(0.725)	83.6(0.793)
	MSE	83.8(1.69)	58.5(1.55)	36.4(1.54)	7.29(0.425)	64.4(1.74)	48.1(1.43)	35.5(3.24)
	MAE	4.11(0.125)	35.6(2.18)	20.3(2.95)	6.56(0.394)	39.5(2.05)	24.5(1.85)	18.9(1.57)
CIFAR100	ALDR-KL	73.3(0.545)	**65.6(0.435)**	49.0(0.581)	7.04(0.566)	**70.2(0.503)**	**63.7(0.909)**	55.8(0.718)
	LDR-KL	**73.5(0.835)**	65.2(0.623)	**49.1(0.759)**	7.09(0.673)	69.8(0.649)	63.4(0.647)	56.6(0.791)
	CE	72.5(0.46)	62.3(0.622)	44.5(1.31)	6.98(0.658)	69.3(0.36)	62.0(0.399)	56.7(0.256)
	SCE	72.5(0.891)	62.7(0.833)	43.2(0.764)	7.41(0.26)	68.8(0.291)	62.5(0.761)	56.7(0.669)
	GCE	73.0(0.409)	62.5(1.01)	44.4(0.824)	7.52(0.318)	68.9(0.559)	62.8(0.613)	**56.8(0.51)**
	TGCE	72.4(0.447)	63.5(0.568)	41.2(0.437)	7.46(0.358)	69.5(0.342)	64.1(0.358)	55.4(0.447)
	WW	65.8(0.454)	51.4(0.83)	32.8(0.992)	7.65(0.448)	63.7(0.408)	60.0(0.319)	49.7(0.616)
	JS	71.4(0.534)	65.0(0.837)	40.1(1.44)	7.44(0.336)	70.0(0.362)	61.8(1.02)	53.2(0.684)
	CS	16.9(0.687)	8.32(0.821)	2.93(0.252)	2.0(0.008)	13.2(0.251)	8.94(0.539)	5.86(0.63)
	RLL	50.7(2.22)	43.1(1.08)	23.6(0.998)	6.9(0.751)	48.4(1.02)	42.4(2.53)	28.1(1.66)
	NCE+RCE	14.9(1.15)	10.1(0.349)	8.23(0.308)	3.6(1.18)	13.7(1.34)	9.88(0.57)	6.15(0.388)
	NCE+AUL	15.4(0.516)	12.4(0.327)	9.48(0.264)	5.09(1.52)	14.1(0.742)	11.9(0.431)	8.02(0.514)
	NCE+AGCE	15.1(0.399)	12.3(0.494)	9.52(0.492)	4.96(1.71)	14.1(0.414)	11.9(0.596)	8.04(0.371)
	MSE	22.1(0.783)	16.4(0.64)	13.6(0.382)	**7.95(0.614)**	18.7(0.426)	15.2(0.316)	10.3(0.728)
	MAE	2.13(0.175)	2.13(0.174)	1.77(0.284)	2.01(0.205)	2.13(0.175)	2.07(0.087)	1.93(0.378)
Tiny-ImageNet	ALDR-KL	59.5(0.166)	**49.9(0.199)**	**32.7(0.358)**	2.13(0.681)	55.9(0.295)	49.2(0.557)	42.5(0.248)
	LDR-KL	59.1(0.443)	49.6(0.361)	32.6(0.255)	2.21(0.702)	55.6(0.222)	48.9(0.27)	**43.8(0.21)**
	CE	**59.6(0.258)**	47.8(0.347)	22.8(0.435)	2.1(0.356)	57.1(0.3)	**51.7(0.439)**	43.3(0.31)
	SCE	59.4(0.364)	47.5(0.566)	22.3(0.766)	**2.75(0.925)**	57.0(0.288)	**51.7(0.456)**	43.2(0.301)
	GCE	59.5(0.364)	47.3(0.53)	19.6(0.816)	2.19(0.133)	**57.2(0.276)**	**51.7(0.263)**	42.9(0.32)
	TGCE	57.8(0.215)	43.0(0.394)	17.7(0.783)	1.85(0.301)	54.9(0.253)	49.4(0.281)	38.6(0.503)
	WW	44.4(0.49)	31.8(0.349)	17.3(0.502)	1.98(0.199)	41.3(0.214)	37.4(0.273)	26.2(0.408)
	JS	49.9(0.386)	28.3(0.797)	7.49(0.93)	2.55(0.349)	44.8(0.538)	30.7(0.292)	19.7(1.37)
	CS	1.86(0.209)	1.28(0.135)	1.0(0.0)	0.988(0.024)	1.47(0.14)	1.32(0.087)	1.14(0.141)
	RLL	13.2(1.22)	5.82(0.359)	4.33(0.254)	1.96(0.603)	9.55(0.64)	5.66(0.352)	4.09(0.178)
	NCE+RCE	2.96(0.125)	2.93(0.105)	2.35(0.191)	1.28(0.18)	2.73(0.243)	2.59(0.25)	2.18(0.243)
	NCE+AUL	3.42(0.135)	3.07(0.167)	2.75(0.361)	1.43(0.37)	3.26(0.258)	3.13(0.171)	2.41(0.206)
	NCE+AGCE	3.44(0.28)	3.24(0.155)	2.66(0.31)	1.47(0.407)	3.49(0.194)	3.1(0.161)	2.39(0.229)
	MSE	4.45(0.162)	1.99(0.215)	1.95(0.202)	1.47(0.154)	2.18(0.231)	1.99(0.173)	1.84(0.167)
	MAE	1.06(0.051)	1.82(0.183)	1.43(0.176)	1.41(0.226)	1.77(0.186)	1.75(0.175)	1.7(0.239)

Table 10. Testing Top-3 Accuracy (%) with mean and standard deviation for tabular dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Vowel	ALDR-KL	**97.2(0.404)**	**89.9(2.02)**	**71.9(5.73)**	**33.9(8.04)**	**97.8(1.34)**	**97.8(1.96)**	**93.5(1.51)**
	LDR-KL	**97.2(1.18)**	88.3(2.08)	63.4(4.44)	32.9(7.58)	96.4(1.21)	96.6(1.03)	92.9(2.12)
	CE	94.7(2.25)	89.3(2.9)	72.1(4.12)	33.7(3.04)	97.4(0.495)	94.7(2.34)	91.3(1.03)
	SCE	96.6(1.03)	89.7(1.62)	69.5(5.21)	29.5(4.62)	95.6(1.87)	95.0(1.69)	90.9(2.63)
	GCE	95.0(1.43)	88.9(2.12)	73.3(2.36)	28.7(5.44)	94.3(1.87)	95.6(1.03)	90.5(1.37)
	TGCE	95.4(1.87)	87.1(3.75)	73.9(2.16)	28.7(5.44)	94.3(1.87)	95.6(1.03)	90.5(1.37)
	WW	95.8(2.06)	88.5(0.495)	68.9(7.27)	**33.9(4.07)**	95.6(1.51)	93.3(2.97)	91.1(2.06)
	JS	93.3(1.98)	89.5(2.27)	73.3(3.92)	31.9(3.76)	93.7(0.756)	91.7(2.96)	89.3(1.76)
	CS	95.8(0.756)	85.7(3.69)	60.6(6.61)	31.1(5.25)	94.3(3.53)	92.3(1.64)	87.3(2.44)
	RLL	90.7(1.96)	88.1(1.62)	73.5(1.62)	27.3(6.09)	90.7(1.49)	90.5(2.44)	87.3(0.808)
	NCE+RCE	86.3(1.37)	82.0(2.06)	57.4(7.27)	29.3(3.0)	88.5(2.27)	88.3(2.08)	85.3(1.76)
	NCE+AUL	84.2(1.21)	81.2(2.68)	55.6(7.34)	32.1(3.28)	86.5(1.03)	87.1(1.96)	85.3(1.64)
	NCE+AGCE	83.8(1.81)	81.6(2.96)	53.7(6.11)	26.1(5.25)	86.9(0.903)	87.5(1.98)	84.8(1.81)
	MSE	86.7(3.34)	81.6(2.16)	58.8(4.62)	29.1(5.97)	86.9(1.43)	83.0(1.49)	81.4(2.44)
	MAE	29.3(2.21)	30.5(3.96)	29.1(2.59)	33.3(4.19)	29.9(4.59)	30.5(4.67)	30.7(4.85)
Letter	ALDR-KL	**90.1(0.366)**	**88.1(0.264)**	84.7(0.182)	**60.5(2.37)**	**90.3(0.471)**	**89.0(0.276)**	**86.3(0.293)**
	LDR-KL	89.3(0.136)	**88.1(0.443)**	**84.9(0.481)**	59.8(1.9)	88.8(0.208)	88.4(0.35)	85.5(0.279)
	CE	87.7(0.22)	85.1(0.246)	79.6(0.817)	32.0(3.8)	87.1(0.185)	85.9(0.36)	81.2(0.297)
	SCE	87.5(0.34)	85.1(0.211)	79.0(0.997)	32.0(3.56)	87.1(0.191)	85.7(0.225)	80.8(0.802)
	GCE	87.3(0.294)	84.6(0.287)	77.9(0.858)	29.4(4.31)	86.9(0.28)	85.6(0.288)	80.5(0.329)
	TGCE	87.2(0.186)	84.6(0.217)	78.8(1.35)	29.5(4.42)	86.9(0.33)	85.5(0.217)	80.6(0.215)
	WW	86.8(0.297)	85.4(0.45)	81.8(1.0)	32.6(1.77)	85.5(0.256)	83.6(0.134)	78.5(1.01)
	JS	86.0(0.402)	83.4(0.22)	71.5(1.03)	21.7(3.88)	85.9(0.22)	84.4(0.555)	78.3(0.83)
	CS	86.7(0.301)	84.3(0.25)	81.1(0.331)	47.2(2.34)	86.0(0.173)	83.9(0.439)	78.6(0.48)
	RLL	83.6(0.236)	76.9(0.927)	49.3(1.38)	11.8(2.55)	83.1(0.34)	81.5(0.499)	72.0(0.661)
	NCE+RCE	70.2(1.46)	64.1(2.54)	55.3(5.12)	23.6(3.49)	67.1(2.24)	52.9(2.01)	48.6(1.68)
	NCE+AUL	69.0(3.48)	62.4(1.84)	53.1(3.17)	18.3(2.56)	62.6(1.67)	52.9(3.25)	45.7(1.75)
	NCE+AGCE	61.1(1.29)	55.2(3.29)	42.9(4.15)	15.1(2.81)	57.4(2.64)	46.3(2.72)	41.7(3.57)
	MSE	38.9(1.24)	34.4(3.15)	25.4(2.67)	14.3(1.79)	36.2(2.18)	34.3(1.97)	32.3(2.21)
	MAE	13.2(1.31)	13.2(1.38)	12.7(1.3)	11.7(2.27)	13.1(1.49)	13.1(1.48)	13.1(1.64)
News20	ALDR-KL	**97.8(0.145)**	**95.8(0.149)**	**89.4(0.908)**	**40.9(3.67)**	97.3(0.157)	96.8(0.222)	96.1(0.154)
	LDR-KL	**97.8(0.213)**	95.8(0.137)	89.5(1.26)	40.1(6.17)	**97.5(0.109)**	**96.9(0.106)**	**96.4(0.286)**
	CE	86.9(0.38)	57.8(2.09)	31.0(2.37)	14.4(1.14)	86.0(0.375)	86.0(0.607)	85.6(0.287)
	SCE	85.0(0.469)	53.3(1.21)	30.5(2.92)	15.9(2.44)	84.1(0.653)	84.4(0.931)	84.1(0.866)
	GCE	77.8(0.963)	45.1(0.789)	28.8(1.98)	15.3(0.847)	76.5(1.2)	78.6(1.31)	78.5(0.873)
	TGCE	76.1(1.08)	44.4(1.23)	28.6(1.82)	15.3(0.847)	74.7(1.53)	77.2(1.39)	77.0(1.05)
	WW	80.2(1.05)	57.5(3.19)	38.0(2.64)	15.6(2.52)	79.6(1.15)	80.7(1.14)	79.5(0.95)
	JS	43.6(0.792)	28.4(1.74)	21.8(1.84)	15.6(1.19)	44.3(1.03)	40.3(0.734)	39.7(1.08)
	CS	92.0(0.925)	92.6(0.76)	76.7(1.9)	19.8(2.28)	91.9(1.05)	90.8(1.02)	89.8(0.652)
	RLL	25.5(0.168)	20.9(1.7)	19.1(2.88)	15.6(1.22)	26.1(0.421)	26.1(0.729)	25.9(0.374)
	NCE+RCE	95.6(0.078)	82.3(3.0)	35.0(2.7)	15.0(1.33)	93.8(2.26)	71.9(1.44)	67.6(0.56)
	NCE+AUL	65.7(2.07)	32.2(1.75)	24.1(1.82)	15.5(0.745)	56.1(1.93)	49.9(2.48)	47.4(1.91)
	NCE+AGCE	41.7(3.43)	23.5(2.23)	19.8(2.38)	16.5(0.775)	37.0(0.911)	31.6(2.49)	31.6(1.61)
	MSE	21.9(2.69)	20.5(1.53)	17.7(1.11)	14.8(0.709)	20.6(1.71)	19.3(1.04)	19.4(2.13)
	MAE	16.0(0.316)	15.6(0.867)	15.4(0.741)	15.4(0.685)	15.8(0.525)	16.0(0.328)	16.0(0.328)
ALOI	ALDR-KL	**98.4(0.099)**	**95.3(0.122)**	90.7(0.108)	**45.1(0.612)**	**97.2(0.06)**	**94.7(0.117)**	**91.3(0.161)**
	LDR-KL	97.5(0.073)	**95.3(0.132)**	**91.2(0.11)**	**45.1(0.749)**	95.8(0.546)	93.6(0.051)	89.3(0.075)
	CE	86.6(0.172)	80.9(0.221)	60.6(0.424)	8.2(0.455)	85.4(0.223)	78.4(0.21)	52.6(0.402)
	SCE	86.1(0.147)	79.8(0.266)	58.5(0.616)	8.19(0.505)	84.7(0.217)	77.3(0.178)	50.4(0.273)
	GCE	83.9(0.162)	74.8(0.257)	48.2(0.306)	6.12(0.47)	82.0(0.268)	71.6(0.326)	41.5(0.256)
	TGCE	83.1(0.256)	73.6(0.266)	47.1(0.484)	6.12(0.47)	81.1(0.152)	70.3(0.195)	39.8(0.36)
	WW	77.5(0.266)	66.7(0.619)	46.3(0.244)	6.94(0.228)	67.0(0.503)	45.1(0.392)	22.4(0.373)
	JS	7.68(0.511)	6.84(0.488)	3.73(0.544)	0.491(0.175)	7.46(0.684)	5.62(0.413)	3.66(0.377)
	CS	91.7(0.182)	83.8(0.234)	42.4(0.427)	3.63(0.131)	90.0(0.069)	86.2(0.123)	71.1(0.228)
	RLL	1.54(0.335)	1.03(0.231)	0.574(0.147)	0.356(0.079)	1.39(0.315)	1.0(0.192)	0.82(0.106)
	NCE+RCE	5.93(0.518)	5.63(0.498)	4.64(0.508)	0.7(0.316)	6.1(0.1)	4.68(0.195)	3.23(0.231)
	NCE+AUL	6.23(0.554)	4.18(0.349)	1.92(0.318)	0.354(0.108)	5.54(0.72)	4.0(0.304)	2.63(0.32)
	NCE+AGCE	4.93(0.448)	2.83(0.361)	1.3(0.212)	0.441(0.139)	4.19(0.666)	2.98(0.436)	1.92(0.244)
	MSE	0.669(0.202)	0.528(0.091)	0.463(0.095)	0.343(0.061)	0.613(0.141)	0.53(0.06)	0.467(0.05)
	MAE	0.441(0.067)	0.441(0.067)	0.363(0.099)	0.311(0.067)	0.441(0.067)	0.42(0.07)	0.396(0.082)

Table 11. Testing Top-3 Accuracy (%) with mean and standard deviation for image dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Kuzushiji-49	ALDR-KL	98.6(0.147)	**97.8(0.094)**	95.2(0.551)	**60.8(1.21)**	98.1(0.053)	97.8(0.062)	96.9(0.187)
	LDR-KL	98.6(0.048)	97.7(0.071)	**95.7(0.729)**	60.2(1.29)	98.1(0.063)	97.7(0.057)	96.9(0.181)
	CE	98.8(0.074)	97.7(0.116)	91.0(0.304)	47.0(2.47)	98.1(0.059)	97.7(0.068)	97.1(0.15)
	SCE	98.9(0.113)	97.7(0.172)	90.5(0.534)	46.1(2.06)	98.1(0.042)	97.7(0.025)	97.0(0.111)
	GCE	98.8(0.142)	97.8(0.09)	90.6(0.285)	44.7(1.8)	98.0(0.045)	97.7(0.05)	97.0(0.114)
	TGCE	98.7(0.126)	97.8(0.102)	90.5(0.216)	45.6(1.1)	98.0(0.044)	97.7(0.064)	97.0(0.135)
	WW	**99.0(0.105)**	97.1(0.092)	91.5(0.487)	45.2(1.23)	**98.2(0.066)**	**98.1(0.081)**	**97.7(0.108)**
	JS	98.7(0.066)	95.3(0.615)	91.0(0.845)	45.0(0.9)	97.0(0.911)	95.4(0.101)	94.6(0.097)
	CS	98.6(0.066)	92.3(2.0)	12.3(3.35)	6.21(0.246)	98.1(0.039)	95.2(0.708)	66.3(2.87)
	RLL	93.0(0.048)	92.6(0.101)	84.1(1.25)	42.8(0.771)	92.7(0.045)	90.5(0.84)	91.4(0.834)
	NCE+RCE	92.5(0.146)	92.1(0.142)	71.8(0.931)	16.7(1.33)	91.4(0.328)	89.4(0.181)	86.8(0.801)
	NCE+AUL	92.6(0.062)	87.2(0.861)	68.2(1.17)	18.3(1.54)	92.4(0.137)	90.8(0.101)	90.4(0.692)
	NCE+AGCE	92.4(0.029)	79.4(0.778)	59.6(0.927)	18.1(1.46)	92.1(0.091)	89.8(0.431)	89.3(0.175)
	MSE	84.1(1.69)	58.7(1.56)	36.3(1.8)	10.8(0.804)	64.5(1.75)	48.9(1.44)	35.9(3.28)
	MAE	6.12(0.0)	35.5(2.36)	20.3(2.91)	9.9(0.879)	39.6(2.09)	24.5(1.81)	18.5(1.6)
CIFAR100	ALDR-KL	79.7(0.586)	**72.6(0.306)**	**56.8(0.778)**	10.0(0.895)	**76.6(0.749)**	**70.7(0.756)**	63.6(0.902)
	LDR-KL	**80.1(0.397)**	72.1(0.534)	56.6(0.607)	10.2(0.824)	76.3(0.586)	70.3(0.463)	**64.2(0.641)**
	CE	79.1(0.492)	69.5(0.637)	52.5(1.06)	10.5(0.695)	75.5(0.318)	69.7(0.636)	64.1(0.289)
	SCE	79.0(0.779)	69.8(0.817)	51.4(0.947)	**11.1(1.05)**	75.2(0.298)	70.1(0.499)	63.9(0.671)
	GCE	79.5(0.318)	69.5(1.39)	52.3(0.677)	10.1(1.11)	75.4(0.726)	70.0(0.392)	64.1(0.345)
	TGCE	78.8(0.469)	70.4(0.846)	49.5(0.466)	10.1(1.02)	75.9(0.324)	**70.7(0.56)**	62.8(0.582)
	WW	74.6(0.441)	60.8(0.6)	41.1(0.586)	10.7(0.738)	72.2(0.456)	68.1(0.296)	58.7(0.646)
	JS	77.8(0.405)	71.8(0.735)	46.4(1.27)	10.2(0.821)	76.5(0.395)	68.4(0.778)	59.8(0.969)
	CS	21.9(0.744)	11.6(1.04)	3.97(0.407)	3.0(0.004)	17.6(0.812)	12.0(0.547)	8.52(0.847)
	RLL	54.9(2.19)	47.3(1.24)	26.9(0.785)	10.1(0.527)	52.5(1.25)	46.3(2.71)	31.8(1.78)
	NCE+RCE	17.0(0.403)	14.0(0.664)	11.4(0.654)	5.64(1.38)	15.7(0.471)	13.1(0.71)	8.5(0.457)
	NCE+AUL	20.0(0.521)	16.6(0.49)	13.1(0.448)	7.76(0.888)	18.7(0.509)	16.3(0.669)	11.4(0.316)
	NCE+AGCE	19.8(0.791)	16.6(0.587)	13.2(0.814)	7.61(0.501)	18.6(0.602)	16.2(0.847)	11.4(0.333)
	MSE	25.4(0.644)	21.1(0.755)	18.3(0.542)	10.9(0.766)	23.4(0.443)	19.9(0.563)	14.4(0.523)
	MAE	3.14(0.21)	3.13(0.208)	2.93(0.22)	2.94(0.373)	3.11(0.171)	2.97(0.251)	3.0(0.308)
Tiny-ImageNet	ALDR-KL	**65.9(0.139)**	**56.5(0.28)**	38.7(0.249)	3.28(0.78)	61.9(0.237)	55.7(0.669)	49.0(0.214)
	LDR-KL	65.2(0.331)	56.1(0.515)	**38.8(0.456)**	3.01(0.754)	61.5(0.148)	55.4(0.454)	**50.4(0.217)**
	CE	**65.9(0.149)**	54.4(0.565)	29.0(0.584)	3.05(0.38)	**63.5(0.214)**	58.3(0.501)	49.9(0.294)
	SCE	**65.9(0.366)**	54.2(0.389)	28.5(0.937)	**3.79(0.769)**	63.3(0.482)	**58.0(0.267)**	49.8(0.166)
	GCE	65.7(0.219)	54.1(0.557)	25.2(0.874)	2.99(0.473)	63.4(0.168)	57.8(0.278)	49.5(0.193)
	TGCE	64.2(0.129)	49.9(0.129)	23.0(0.792)	3.22(0.22)	61.3(0.243)	55.5(0.206)	44.9(0.433)
	WW	52.4(0.338)	39.3(0.292)	22.9(0.558)	2.66(0.283)	49.7(0.393)	45.6(0.424)	33.4(0.192)
	JS	55.2(0.546)	32.2(0.671)	10.5(0.703)	3.61(0.401)	49.8(0.616)	34.8(0.222)	22.6(1.17)
	CS	2.57(0.243)	1.67(0.123)	1.5(0.0)	1.5(0.0)	2.07(0.141)	1.7(0.137)	1.52(0.089)
	RLL	14.2(1.26)	7.96(0.586)	5.92(0.363)	3.22(0.26)	10.7(0.59)	7.41(0.588)	5.44(0.067)
	NCE+RCE	4.21(0.207)	4.09(0.173)	3.36(0.124)	1.82(0.281)	4.01(0.345)	3.72(0.257)	3.1(0.227)
	NCE+AUL	4.79(0.114)	4.39(0.105)	3.68(0.296)	2.13(0.39)	4.63(0.22)	4.36(0.365)	3.59(0.375)
	NCE+AGCE	4.87(0.252)	4.5(0.188)	3.9(0.262)	2.18(0.417)	4.79(0.281)	4.45(0.268)	3.55(0.174)
	MSE	6.1(0.247)	2.93(0.161)	2.75(0.298)	2.01(0.361)	3.01(0.291)	2.89(0.188)	2.78(0.197)
	MAE	1.57(0.063)	2.51(0.254)	2.06(0.254)	2.11(0.243)	2.49(0.205)	2.51(0.185)	2.47(0.189)

Table 12. Testing Top-4 Accuracy (%) with mean and standard deviation for tabular dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Vowel	ALDR-KL	**98.8(0.99)**	91.7(0.99)	79.8(3.06)	37.8(6.25)	**98.8(0.756)**	**99.0(1.11)**	**98.2(0.756)**
	LDR-KL	97.6(0.808)	92.1(2.34)	73.5(6.8)	39.2(4.11)	97.8(1.18)	97.6(1.51)	96.8(2.96)
	CE	97.8(0.756)	93.3(1.87)	78.2(2.9)	36.6(6.56)	98.2(1.74)	97.8(0.756)	95.8(0.756)
	SCE	97.6(0.808)	93.3(1.64)	75.6(6.86)	37.2(4.84)	98.2(0.756)	97.4(1.51)	94.7(1.18)
	GCE	97.8(0.756)	**93.9(1.69)**	80.6(2.74)	37.2(5.32)	97.6(1.03)	97.8(0.404)	95.4(1.21)
	TGCE	97.8(0.756)	93.1(1.49)	80.0(3.34)	37.2(5.32)	98.0(0.639)	97.8(0.404)	95.2(1.18)
	WW	98.0(0.0)	93.3(1.37)	78.2(3.65)	39.2(7.04)	97.2(0.99)	97.6(1.37)	96.4(1.03)
	JS	96.0(1.28)	93.5(0.808)	79.2(3.65)	37.4(6.64)	97.2(1.34)	96.4(1.21)	93.3(1.87)
	CS	98.0(1.43)	90.5(3.48)	72.7(4.19)	39.4(6.03)	98.2(0.404)	96.8(1.96)	93.1(2.25)
	RLL	94.1(1.18)	90.9(1.11)	**82.0(0.404)**	34.9(4.36)	95.6(1.03)	94.1(0.756)	90.5(0.808)
	NCE+RCE	92.7(1.18)	89.1(2.67)	74.1(2.44)	**40.2(2.81)**	92.9(0.639)	91.7(1.18)	91.1(1.18)
	NCE+AUL	91.3(1.21)	88.5(2.97)	70.1(5.91)	36.2(6.21)	93.7(0.756)	93.5(1.03)	91.7(0.756)
	NCE+AGCE	91.9(2.63)	88.1(2.74)	68.9(4.62)	38.6(5.09)	93.1(0.99)	93.1(1.18)	91.5(0.495)
	MSE	90.5(0.808)	89.5(1.37)	72.3(1.87)	36.2(5.58)	91.5(2.36)	90.9(0.639)	89.9(1.11)
	MAE	41.6(5.69)	42.8(6.34)	44.6(5.47)	44.2(5.17)	39.0(2.9)	38.8(2.68)	40.4(4.04)
Letter	ALDR-KL	**92.4(0.183)**	**90.8(0.268)**	**87.9(1.07)**	**65.9(3.19)**	**92.6(0.275)**	**91.5(0.37)**	**89.7(0.411)**
	LDR-KL	91.7(0.173)	90.7(0.528)	87.5(0.541)	64.0(4.31)	91.2(0.434)	91.0(0.437)	89.1(0.252)
	CE	90.1(0.195)	88.5(0.265)	83.1(1.81)	40.9(1.91)	89.4(0.203)	88.1(0.256)	85.6(0.344)
	SCE	90.0(0.185)	88.4(0.369)	82.8(1.94)	37.6(3.88)	89.4(0.133)	88.2(0.229)	85.5(0.483)
	GCE	89.7(0.146)	87.9(0.138)	81.8(2.05)	32.9(4.88)	89.2(0.243)	88.0(0.288)	85.3(0.377)
	TGCE	89.5(0.105)	87.9(0.138)	83.6(1.43)	32.8(5.02)	89.2(0.3)	87.9(0.193)	85.2(0.386)
	WW	88.9(0.208)	88.4(0.202)	85.6(0.614)	38.5(2.85)	88.3(0.1)	86.8(0.126)	83.7(0.084)
	JS	88.4(0.314)	86.9(0.121)	76.4(0.812)	25.9(3.73)	88.1(0.287)	87.1(0.402)	84.1(0.213)
	CS	89.0(0.2)	87.6(0.419)	84.6(0.934)	54.3(4.58)	88.5(0.3)	86.9(0.227)	84.0(0.169)
	RLL	86.1(0.086)	80.6(0.394)	56.9(0.491)	16.5(2.56)	85.8(0.317)	84.5(0.229)	79.2(1.3)
	NCE+RCE	71.3(1.48)	65.1(2.36)	56.2(5.03)	26.6(3.99)	68.1(2.29)	55.5(0.956)	49.5(1.72)
	NCE+AUL	70.2(3.73)	63.5(1.89)	54.5(3.35)	25.0(2.62)	63.7(1.8)	53.9(3.16)	46.8(2.77)
	NCE+AGCE	62.0(1.41)	55.9(3.61)	44.7(3.4)	19.7(1.44)	58.4(2.71)	48.5(1.86)	44.7(1.55)
	MSE	46.5(1.16)	41.1(2.45)	31.7(2.62)	18.2(1.89)	42.9(2.26)	41.4(2.01)	39.9(1.89)
	MAE	16.9(1.49)	17.1(1.55)	16.8(1.84)	15.2(2.19)	16.9(1.47)	17.0(1.51)	16.8(1.36)
News20	ALDR-KL	**98.3(0.144)**	**96.6(0.203)**	91.4(0.972)	44.5(2.68)	**98.2(0.121)**	**97.8(0.216)**	97.4(0.326)
	LDR-KL	98.3(0.204)	96.5(0.221)	**91.6(1.04)**	**45.5(4.8)**	98.2(0.172)	97.7(0.136)	**97.7(0.162)**
	CE	90.3(0.434)	67.1(1.18)	38.7(3.68)	19.9(1.02)	89.7(0.264)	89.8(0.34)	89.5(0.086)
	SCE	88.9(0.317)	63.2(0.651)	37.2(4.88)	19.9(0.539)	88.2(0.539)	88.8(0.601)	88.4(0.375)
	GCE	83.7(0.62)	55.8(0.96)	36.0(3.91)	20.1(0.779)	82.6(1.16)	83.5(0.744)	83.5(0.835)
	TGCE	81.9(0.728)	55.3(0.437)	36.0(3.91)	20.1(0.779)	80.6(1.36)	82.2(1.6)	82.2(1.35)
	WW	85.4(0.899)	64.9(2.17)	42.5(3.36)	20.8(2.09)	84.9(0.685)	85.6(0.771)	84.9(0.595)
	JS	51.7(0.279)	36.2(0.842)	27.7(0.745)	20.6(1.31)	51.4(1.31)	49.3(0.713)	48.7(1.12)
	CS	95.2(0.543)	95.2(0.333)	81.5(0.953)	23.8(2.6)	95.0(0.726)	95.2(0.606)	94.1(0.646)
	RLL	33.2(0.464)	26.3(1.52)	25.3(3.24)	20.2(1.14)	33.8(0.722)	34.2(0.574)	34.4(0.561)
	NCE+RCE	96.3(0.232)	84.5(3.1)	39.7(3.62)	20.5(0.449)	94.6(2.29)	72.6(1.57)	68.1(0.657)
	NCE+AUL	66.4(1.63)	35.4(0.569)	31.4(1.38)	20.5(0.75)	56.4(1.76)	50.5(2.16)	47.7(2.05)
	NCE+AGCE	42.7(3.07)	29.0(1.18)	27.2(3.88)	22.0(1.82)	38.8(3.08)	33.2(1.85)	33.8(1.6)
	MSE	28.2(2.11)	25.0(0.965)	23.7(1.23)	19.8(0.615)	26.7(2.43)	26.2(2.06)	25.8(1.35)
	MAE	20.8(0.453)	20.9(0.462)	20.2(1.03)	19.8(1.34)	20.8(0.453)	20.8(0.453)	20.8(0.453)
ALOI	ALDR-KL	**98.8(0.073)**	**96.2(0.118)**	92.0(0.158)	48.1(0.334)	**97.8(0.033)**	**95.8(0.107)**	**93.3(0.166)**
	LDR-KL	98.1(0.096)	96.1(0.122)	**92.4(0.118)**	**48.4(0.474)**	96.3(0.382)	94.7(0.076)	91.6(0.167)
	CE	88.5(0.072)	83.5(0.196)	65.5(0.455)	9.82(0.444)	87.6(0.137)	81.6(0.095)	58.2(0.527)
	SCE	88.1(0.11)	82.8(0.253)	63.4(0.497)	9.89(0.522)	86.9(0.143)	80.7(0.132)	56.3(0.456)
	GCE	86.2(0.186)	78.2(0.15)	53.7(0.392)	7.58(0.509)	84.7(0.222)	75.8(0.209)	47.0(0.241)
	TGCE	85.5(0.144)	77.3(0.288)	52.4(0.604)	7.58(0.509)	83.9(0.169)	74.4(0.36)	45.1(0.302)
	WW	81.6(0.328)	72.1(0.618)	52.6(0.196)	8.81(0.195)	73.0(0.4)	51.7(0.272)	27.4(0.323)
	JS	9.39(0.643)	8.38(0.517)	4.67(0.542)	0.593(0.211)	9.21(0.77)	6.77(0.627)	4.51(0.314)
	CS	93.1(0.146)	85.8(0.327)	45.4(0.53)	4.27(0.23)	91.4(0.046)	88.1(0.149)	75.5(0.179)
	RLL	1.88(0.336)	1.28(0.276)	0.763(0.169)	0.474(0.126)	1.71(0.3)	1.33(0.135)	1.06(0.063)
	NCE+RCE	7.19(0.531)	6.96(0.552)	5.92(0.604)	0.833(0.397)	7.32(0.197)	5.82(0.175)	4.02(0.167)
	NCE+AUL	7.59(0.566)	5.26(0.362)	2.32(0.329)	0.554(0.208)	6.93(0.714)	4.96(0.422)	3.36(0.199)
	NCE+AGCE	5.97(0.51)	3.52(0.487)	1.59(0.277)	0.598(0.182)	5.22(0.593)	3.7(0.46)	2.54(0.151)
	MSE	0.806(0.206)	0.683(0.13)	0.646(0.125)	0.43(0.06)	0.73(0.187)	0.693(0.154)	0.643(0.137)
	MAE	0.543(0.061)	0.552(0.056)	0.5(0.105)	0.48(0.081)	0.552(0.056)	0.541(0.049)	0.498(0.074)

Table 13. Testing Top-4 Accuracy (%) with mean and standard deviation for image dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Kuzushiji-49	ALDR-KL	98.8(0.15)	**98.2(0.059)**	**95.8(0.238)**	**64.1(1.81)**	98.4(0.071)	98.2(0.039)	97.9(0.038)
	LDR-KL	98.8(0.035)	97.9(0.314)	95.7(0.862)	63.7(0.856)	98.4(0.029)	98.1(0.04)	97.9(0.097)
	CE	99.1(0.102)	98.0(0.114)	92.3(0.062)	51.0(2.52)	98.4(0.044)	98.1(0.08)	97.8(0.062)
	SCE	99.0(0.123)	98.1(0.077)	91.7(0.734)	50.1(2.15)	98.4(0.049)	98.1(0.029)	97.8(0.061)
	GCE	99.0(0.138)	98.2(0.05)	91.5(0.295)	48.9(1.81)	98.4(0.021)	98.0(0.041)	97.8(0.052)
	TGCE	99.0(0.143)	97.9(0.321)	91.4(0.28)	49.9(0.791)	98.4(0.035)	98.0(0.058)	97.8(0.101)
	WW	**99.1(0.138)**	97.6(0.108)	93.5(0.487)	50.0(1.34)	**98.5(0.096)**	**98.4(0.065)**	**98.4(0.12)**
	JS	98.8(0.019)	95.8(0.574)	91.4(0.902)	48.8(1.06)	97.4(0.894)	95.8(0.064)	95.4(0.071)
	CS	98.7(0.05)	93.4(1.88)	14.5(3.48)	8.33(0.264)	98.2(0.043)	96.2(0.791)	70.3(2.96)
	RLL	93.1(0.033)	92.8(0.049)	84.6(1.21)	46.3(0.693)	92.9(0.042)	91.3(0.667)	91.7(1.14)
	NCE+RCE	92.8(0.244)	92.4(0.157)	72.1(0.951)	20.9(1.66)	92.3(0.172)	91.5(0.024)	89.2(0.39)
	NCE+AUL	92.8(0.041)	87.4(0.879)	68.6(0.937)	23.5(2.27)	92.7(0.036)	93.4(0.248)	92.6(0.633)
	NCE+AGCE	92.6(0.051)	78.1(2.97)	59.6(1.44)	22.9(1.55)	92.6(0.058)	93.2(0.262)	91.2(0.213)
	MSE	84.3(1.69)	58.8(1.59)	36.5(1.67)	14.2(0.746)	64.6(1.76)	50.1(1.89)	36.0(3.27)
	MAE	8.19(0.027)	35.5(2.19)	20.6(2.41)	12.8(1.15)	39.7(2.11)	24.6(2.01)	18.7(1.36)
CIFAR100	ALDR-KL	83.7(0.439)	**77.2(0.326)**	**62.2(0.734)**	12.8(0.688)	**80.5(0.613)**	**75.2(0.988)**	69.7(0.54)
	LDR-KL	**83.9(0.248)**	76.7(0.516)	**62.2(0.617)**	13.7(0.831)	80.3(0.297)	74.8(0.822)	**70.1(0.838)**
	CE	82.9(0.64)	74.3(0.716)	58.3(1.11)	13.2(0.785)	79.4(0.356)	74.1(0.661)	69.9(0.247)
	SCE	82.9(0.592)	74.2(0.778)	57.2(1.08)	13.9(1.2)	79.3(0.273)	74.5(0.416)	69.7(0.685)
	GCE	83.2(0.234)	74.8(0.407)	58.1(0.774)	13.0(0.483)	79.4(0.703)	74.5(0.384)	69.9(0.281)
	TGCE	82.6(0.376)	75.1(0.797)	55.3(0.495)	12.4(1.21)	80.0(0.271)	75.1(0.568)	68.5(0.514)
	WW	79.8(0.35)	67.6(0.602)	47.3(0.951)	**13.9(0.91)**	77.3(0.485)	73.4(0.224)	65.7(0.41)
	JS	81.7(0.203)	76.0(1.08)	50.9(1.06)	13.3(0.48)	80.4(0.424)	73.1(1.16)	65.2(0.508)
	CS	26.3(0.856)	14.4(0.907)	4.91(0.68)	4.01(0.07)	21.5(0.882)	15.5(0.83)	11.1(1.09)
	RLL	57.2(2.13)	49.8(1.39)	28.9(0.97)	12.7(0.901)	55.0(1.4)	48.7(3.0)	34.7(1.8)
	NCE+RCE	20.2(0.588)	17.3(1.04)	14.6(0.809)	7.54(1.91)	19.1(0.433)	16.4(0.832)	11.4(0.337)
	NCE+AUL	24.2(0.528)	20.5(0.627)	16.1(0.387)	9.93(1.12)	22.8(0.462)	20.0(0.739)	14.5(0.474)
	NCE+AGCE	24.4(0.816)	20.5(0.623)	16.2(0.535)	10.3(0.729)	22.9(0.578)	20.2(0.981)	14.7(0.516)
	MSE	28.9(0.699)	25.2(0.874)	22.2(0.584)	13.6(0.687)	27.1(0.484)	23.7(0.631)	17.8(0.484)
	MAE	4.16(0.274)	4.14(0.287)	3.87(0.292)	4.04(0.121)	4.14(0.286)	4.07(0.36)	3.81(0.327)
Tiny-ImageNet	ALDR-KL	70.0(0.15)	**61.0(0.381)**	43.6(0.495)	4.73(0.51)	66.0(0.391)	60.5(0.229)	54.3(0.215)
	LDR-KL	69.4(0.174)	60.5(0.45)	**43.8(0.751)**	4.69(0.789)	65.5(0.28)	59.8(0.352)	**55.6(0.331)**
	CE	70.0(0.269)	59.4(0.332)	33.9(0.614)	3.4(0.583)	**67.6(0.208)**	**62.3(0.422)**	55.1(0.149)
	SCE	**70.2(0.192)**	59.0(0.385)	33.3(0.788)	**5.32(0.641)**	67.4(0.448)	62.2(0.052)	55.1(0.262)
	GCE	69.9(0.194)	58.8(0.462)	29.8(0.729)	3.62(0.401)	67.6(0.241)	62.0(0.226)	54.7(0.325)
	TGCE	68.3(0.164)	54.9(0.158)	27.4(0.691)	3.94(0.349)	65.7(0.226)	59.5(0.243)	50.0(0.476)
	WW	58.2(0.329)	45.2(0.408)	27.5(0.388)	3.32(0.356)	55.3(0.436)	51.3(0.48)	39.7(0.376)
	JS	58.9(0.417)	35.0(0.682)	12.9(0.797)	4.65(0.249)	53.3(0.814)	38.0(0.204)	24.9(1.4)
	CS	3.37(0.295)	2.22(0.129)	2.0(0.0)	2.0(0.0)	2.66(0.184)	2.21(0.144)	2.15(0.189)
	RLL	14.8(1.37)	9.59(0.72)	7.23(0.443)	3.59(0.803)	11.5(0.351)	8.95(0.751)	6.64(0.186)
	NCE+RCE	5.5(0.193)	5.36(0.357)	4.24(0.182)	2.38(0.332)	5.28(0.387)	4.89(0.184)	3.93(0.177)
	NCE+AUL	6.27(0.264)	5.6(0.228)	4.7(0.499)	2.71(0.427)	6.0(0.147)	5.58(0.315)	4.72(0.41)
	NCE+AGCE	6.3(0.271)	5.77(0.187)	5.03(0.338)	3.07(0.341)	6.27(0.208)	5.65(0.377)	4.67(0.148)
	MSE	7.64(0.284)	3.7(0.353)	3.58(0.28)	2.56(0.452)	3.68(0.311)	3.69(0.209)	3.62(0.183)
	MAE	2.06(0.113)	3.23(0.305)	2.77(0.342)	2.75(0.419)	3.33(0.328)	3.29(0.166)	3.25(0.207)

Table 14. Testing Top-5 Accuracy (%) with mean and standard deviation for tabular dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Vowel	ALDR-KL	98.6(0.808)	94.7(0.404)	85.1(4.49)	44.2(3.75)	99.2(0.756)	**99.2(0.756)**	**98.8(0.404)**
	LDR-KL	98.6(0.808)	94.9(2.12)	81.0(2.59)	47.1(4.22)	98.8(0.404)	98.8(1.18)	98.0(0.639)
	CE	98.8(1.18)	95.8(1.62)	85.9(3.67)	42.0(5.21)	98.6(0.808)	98.4(0.495)	97.2(0.756)
	SCE	98.8(0.756)	95.4(2.6)	83.2(4.59)	47.7(4.84)	**99.4(0.495)**	98.4(0.495)	96.8(0.404)
	GCE	98.8(0.756)	95.4(2.68)	85.7(3.16)	46.9(4.07)	98.4(0.495)	97.6(1.03)	97.2(0.404)
	TGCE	98.8(0.756)	95.4(2.68)	86.5(3.23)	46.9(4.07)	98.6(0.495)	97.2(1.18)	97.2(0.404)
	WW	98.8(0.756)	**96.0(1.43)**	**86.7(1.74)**	48.3(6.4)	98.2(0.756)	98.8(0.99)	98.2(0.99)
	JS	97.2(2.06)	94.5(1.21)	86.3(2.6)	46.3(5.97)	97.6(1.03)	97.6(0.495)	96.4(1.51)
	CS	**99.4(0.808)**	92.9(1.81)	76.0(4.11)	50.1(8.0)	98.4(1.51)	98.2(1.49)	96.0(0.903)
	RLL	96.0(1.11)	95.4(0.808)	84.4(5.25)	44.2(6.37)	96.4(0.808)	96.4(1.37)	94.5(2.18)
	NCE+RCE	96.4(0.808)	92.9(2.47)	82.2(2.9)	44.8(4.5)	96.2(0.404)	96.0(0.903)	92.5(2.36)
	NCE+AUL	96.2(0.756)	93.3(2.36)	78.6(8.7)	**50.5(4.99)**	96.0(0.639)	95.4(1.21)	93.7(1.18)
	NCE+AGCE	96.2(0.756)	94.3(1.98)	76.8(6.06)	44.6(7.21)	95.6(1.03)	95.4(0.808)	93.5(0.808)
	MSE	93.7(1.34)	92.7(1.96)	75.4(11.9)	47.7(3.09)	94.9(0.639)	94.3(0.808)	93.3(1.03)
	MAE	52.5(5.53)	51.9(5.98)	53.1(4.41)	51.9(5.48)	52.3(5.36)	52.1(5.48)	49.5(4.14)
Letter	ALDR-KL	**94.3(0.336)**	92.9(0.306)	**89.7(0.796)**	70.1(4.21)	**94.3(0.174)**	**93.4(0.229)**	**91.8(0.427)**
	LDR-KL	93.5(0.108)	**93.0(0.406)**	89.5(1.0)	**70.3(3.98)**	92.8(0.468)	92.7(0.216)	91.2(0.399)
	CE	91.8(0.12)	90.6(0.159)	87.7(0.511)	40.5(8.44)	91.6(0.169)	90.1(0.129)	88.6(0.223)
	SCE	91.8(0.143)	90.7(0.169)	87.0(0.584)	43.2(3.21)	91.5(0.117)	90.1(0.314)	88.6(0.379)
	GCE	91.5(0.126)	90.3(0.215)	86.2(0.514)	42.0(2.35)	91.1(0.08)	89.9(0.223)	88.3(0.343)
	TGCE	91.5(0.125)	90.3(0.21)	87.5(0.745)	41.8(2.39)	91.0(0.081)	89.8(0.196)	88.1(0.394)
	WW	91.0(0.108)	90.4(0.159)	88.5(0.6)	42.8(4.43)	90.5(0.116)	89.2(0.153)	86.7(0.223)
	JS	90.3(0.246)	89.2(0.177)	80.7(0.606)	31.8(1.72)	90.1(0.191)	89.3(0.187)	87.1(0.522)
	CS	91.0(0.244)	89.9(0.405)	88.0(0.219)	56.3(5.13)	90.2(0.159)	88.6(0.289)	86.7(0.484)
	RLL	88.4(0.174)	84.0(0.24)	61.6(0.612)	23.6(3.09)	88.0(0.242)	87.0(0.196)	82.8(1.06)
	NCE+RCE	72.1(1.19)	65.8(2.56)	57.1(5.08)	32.8(3.36)	69.6(1.24)	63.9(1.1)	55.6(1.19)
	NCE+AUL	70.9(4.05)	64.3(1.88)	55.9(2.85)	28.4(2.64)	64.5(1.64)	56.0(2.18)	52.0(2.07)
	NCE+AGCE	63.5(2.32)	58.5(3.16)	50.1(3.48)	22.7(1.88)	61.2(2.42)	55.2(1.37)	51.5(1.82)
	MSE	53.4(0.907)	46.9(1.62)	36.5(2.63)	22.2(1.82)	51.0(1.43)	47.4(1.78)	44.9(2.43)
	MAE	21.1(1.16)	21.0(1.3)	20.4(1.5)	20.4(1.75)	21.0(1.34)	20.9(1.5)	20.3(1.31)
News20	ALDR-KL	**98.6(0.072)**	**97.2(0.149)**	**93.0(0.768)**	**50.9(2.19)**	**98.6(0.14)**	98.2(0.157)	98.0(0.337)
	LDR-KL	**98.6(0.132)**	97.0(0.303)	92.9(0.63)	50.2(3.46)	98.5(0.119)	**98.3(0.123)**	**98.1(0.17)**
	CE	92.6(0.306)	74.8(0.619)	45.0(2.67)	25.4(1.15)	91.9(0.17)	92.0(0.254)	91.6(0.226)
	SCE	91.6(0.282)	71.6(0.778)	44.3(4.5)	26.2(1.94)	90.8(0.226)	91.3(0.203)	90.8(0.275)
	GCE	87.6(0.357)	64.3(0.357)	41.3(4.15)	24.7(0.799)	86.6(0.591)	87.0(0.632)	86.9(0.378)
	TGCE	86.2(0.554)	64.0(0.514)	40.9(4.29)	24.7(0.799)	85.5(0.613)	86.2(0.656)	86.2(0.82)
	WW	88.7(0.425)	72.3(1.27)	47.9(3.18)	26.0(2.36)	88.6(0.358)	88.5(0.214)	87.9(0.346)
	JS	58.7(0.538)	44.0(0.663)	34.6(2.3)	25.2(1.55)	58.5(0.951)	57.3(0.561)	57.0(0.977)
	CS	96.7(0.664)	97.2(0.182)	85.1(1.51)	30.0(3.31)	96.7(0.387)	96.9(0.522)	96.4(0.552)
	RLL	40.8(1.05)	33.0(1.57)	31.2(2.23)	25.4(1.18)	40.6(0.837)	41.6(1.03)	41.0(1.17)
	NCE+RCE	96.9(0.201)	86.0(3.13)	42.5(3.52)	25.6(1.96)	95.0(2.02)	73.3(1.42)	68.5(0.604)
	NCE+AUL	67.6(1.11)	41.1(1.0)	37.4(2.68)	24.6(0.774)	56.8(1.36)	51.2(2.35)	48.5(1.9)
	NCE+AGCE	44.6(2.77)	35.0(0.94)	33.7(3.99)	25.0(1.23)	42.4(4.11)	38.7(1.97)	38.8(2.18)
	MSE	33.9(1.96)	30.6(0.669)	29.8(0.907)	24.8(0.62)	32.6(1.89)	31.5(2.33)	31.5(1.61)
	MAE	26.1(0.276)	26.1(0.276)	25.8(1.03)	24.9(1.16)	26.1(0.276)	26.1(0.276)	26.1(0.276)
ALOI	ALDR-KL	**99.0(0.069)**	**96.7(0.118)**	92.8(0.14)	50.7(0.693)	**98.2(0.047)**	**96.5(0.075)**	**94.5(0.187)**
	LDR-KL	98.4(0.075)	**96.7(0.111)**	**93.3(0.146)**	**50.8(0.416)**	97.2(0.34)	95.5(0.068)	92.8(0.147)
	CE	89.8(0.059)	85.5(0.181)	69.1(0.627)	11.4(0.708)	89.1(0.113)	83.7(0.173)	62.5(0.417)
	SCE	89.5(0.1)	84.7(0.166)	67.2(0.647)	11.4(0.615)	88.6(0.141)	83.0(0.181)	60.5(0.569)
	GCE	87.9(0.086)	80.8(0.163)	57.8(0.51)	8.58(0.893)	86.6(0.173)	78.8(0.299)	51.2(0.291)
	TGCE	87.3(0.059)	79.9(0.185)	56.5(0.638)	8.58(0.893)	85.8(0.226)	77.5(0.343)	49.3(0.352)
	WW	84.2(0.41)	76.0(0.52)	57.5(0.175)	10.5(0.205)	77.0(0.362)	57.0(0.344)	31.8(0.303)
	JS	11.0(0.743)	9.69(0.529)	5.55(0.492)	0.62(0.129)	10.6(0.935)	8.18(0.719)	5.32(0.324)
	CS	94.1(0.192)	87.1(0.349)	47.9(0.514)	4.77(0.468)	92.5(0.105)	89.4(0.124)	78.3(0.241)
	RLL	2.17(0.384)	1.53(0.274)	0.87(0.209)	0.524(0.065)	1.98(0.31)	1.57(0.145)	1.27(0.088)
	NCE+RCE	8.55(0.52)	8.34(0.598)	6.95(0.589)	1.03(0.387)	8.44(0.687)	7.11(0.385)	4.73(0.215)
	NCE+AUL	8.76(0.631)	6.27(0.422)	2.65(0.36)	0.702(0.309)	8.07(0.822)	5.79(0.436)	3.98(0.198)
	NCE+AGCE	6.99(0.535)	4.26(0.589)	1.92(0.297)	0.609(0.129)	6.15(0.568)	4.26(0.506)	2.99(0.185)
	MSE	0.926(0.242)	0.811(0.197)	0.744(0.164)	0.493(0.033)	0.88(0.214)	0.83(0.224)	0.8(0.165)
	MAE	0.635(0.067)	0.617(0.089)	0.585(0.092)	0.581(0.1)	0.611(0.09)	0.565(0.108)	0.583(0.088)

Table 15. Testing Top-5 Accuracy (%) with mean and standard deviation for image dataset. The highest values are marked as bold.

Dataset	Loss	Clean	Uniform 0.3	Uniform 0.6	Uniform 0.9	CD 0.1	CD 0.3	CD 0.5
Kuzushiji-49	ALDR-KL	98.9(0.129)	**98.4(0.059)**	**96.5(0.143)**	**67.3(1.97)**	**98.6(0.065)**	98.4(0.037)	98.1(0.151)
	LDR-KL	98.9(0.033)	98.1(0.352)	**96.5(0.571)**	66.4(1.01)	98.5(0.052)	98.4(0.06)	98.2(0.053)
	CE	99.1(0.096)	98.3(0.063)	93.2(0.226)	54.4(1.57)	98.5(0.055)	98.4(0.056)	98.2(0.051)
	SCE	99.1(0.138)	98.1(0.464)	93.0(0.377)	53.0(1.94)	98.6(0.042)	98.4(0.048)	98.1(0.046)
	GCE	99.1(0.148)	**98.4(0.042)**	92.4(0.31)	51.9(2.02)	98.5(0.024)	98.3(0.07)	98.1(0.045)
	TGCE	99.0(0.112)	98.0(0.341)	92.3(0.329)	53.1(0.717)	98.5(0.046)	98.3(0.053)	98.1(0.079)
	WW	**99.2(0.157)**	97.9(0.065)	94.8(0.308)	53.7(0.948)	98.7(0.137)	**98.7(0.1)**	**98.6(0.152)**
	JS	98.9(0.012)	96.0(0.597)	91.7(0.722)	51.9(1.07)	97.9(0.7)	96.1(0.068)	95.8(0.087)
	CS	98.8(0.072)	94.2(1.82)	16.1(3.25)	10.2(0.022)	98.3(0.039)	96.5(0.642)	73.1(2.97)
	RLL	93.2(0.064)	92.9(0.118)	84.8(1.11)	48.9(0.645)	93.1(0.037)	91.7(0.588)	92.4(0.827)
	NCE+RCE	93.0(0.39)	92.4(0.226)	72.5(0.659)	25.0(1.22)	92.7(0.108)	91.9(0.06)	90.7(0.507)
	NCE+AUL	93.0(0.036)	87.6(0.85)	69.0(0.774)	27.3(2.64)	92.9(0.051)	94.0(0.081)	93.7(0.06)
	NCE+AGCE	92.7(0.046)	78.5(2.37)	61.3(0.89)	27.0(1.83)	92.9(0.055)	93.8(0.177)	93.3(0.242)
	MSE	84.4(1.66)	59.0(1.57)	37.3(1.19)	17.3(0.961)	64.8(1.67)	52.0(1.85)	35.8(3.73)
	MAE	10.3(0.069)	35.9(2.23)	21.6(1.27)	15.7(1.07)	39.8(2.21)	24.5(2.2)	19.9(1.39)
CIFAR100	ALDR-KL	86.3(0.455)	**80.3(0.39)**	**66.3(0.585)**	15.7(1.12)	83.1(0.534)	**78.5(0.563)**	73.5(0.581)
	LDR-KL	**86.4(0.376)**	79.7(0.558)	66.2(0.653)	15.9(0.933)	83.0(0.172)	77.8(0.651)	**74.0(0.68)**
	CE	85.5(0.508)	78.0(0.622)	62.2(1.11)	15.8(1.46)	82.2(0.245)	77.4(0.639)	73.6(0.372)
	SCE	85.5(0.628)	77.9(0.97)	61.6(1.06)	16.3(1.57)	82.1(0.264)	77.6(0.507)	73.4(0.516)
	GCE	85.9(0.17)	78.4(0.774)	62.5(0.854)	16.0(1.14)	82.3(0.577)	77.7(0.497)	73.6(0.157)
	TGCE	85.3(0.329)	78.4(0.77)	59.7(0.395)	15.2(1.8)	82.6(0.422)	78.0(0.479)	72.3(0.54)
	WW	83.3(0.409)	72.0(0.522)	52.2(0.918)	**16.8(0.596)**	80.6(0.465)	77.1(0.189)	70.4(0.505)
	JS	84.5(0.349)	79.3(0.8)	54.5(1.04)	15.2(1.04)	**83.4(0.374)**	76.0(1.11)	68.7(0.489)
	CS	29.8(0.888)	16.8(1.1)	6.52(0.596)	5.0(0.0)	24.9(1.1)	18.5(0.597)	13.2(1.12)
	RLL	58.8(2.08)	51.4(1.64)	31.0(0.911)	14.8(0.787)	56.7(1.49)	50.3(3.02)	36.9(1.81)
	NCE+RCE	23.6(0.616)	20.1(1.2)	17.3(0.963)	8.74(2.01)	22.5(0.671)	19.6(0.74)	14.2(0.477)
	NCE+AUL	27.9(0.602)	23.8(0.661)	18.9(0.614)	11.1(3.16)	26.6(0.486)	23.2(0.945)	17.7(0.264)
	NCE+AGCE	28.2(0.841)	24.3(0.492)	19.1(0.85)	11.8(0.931)	26.7(0.48)	23.2(1.08)	17.9(0.182)
	MSE	31.9(0.519)	28.9(0.992)	25.5(0.494)	15.8(0.558)	30.3(0.557)	27.1(0.696)	20.6(0.592)
	MAE	5.11(0.161)	5.12(0.15)	5.15(0.128)	5.03(0.163)	5.03(0.291)	5.13(0.145)	4.88(0.356)
Tiny-ImageNet	ALDR-KL	73.0(0.216)	**64.4(0.354)**	47.4(0.827)	5.5(0.527)	69.0(0.212)	63.8(0.185)	58.1(0.295)
	LDR-KL	72.4(0.174)	63.9(0.477)	**47.8(0.693)**	6.15(0.338)	68.5(0.32)	63.1(0.254)	**58.9(0.383)**
	CE	**73.3(0.219)**	62.9(0.28)	38.0(0.557)	4.36(0.76)	70.8(0.203)	65.5(0.38)	58.7(0.263)
	SCE	**73.3(0.111)**	62.8(0.359)	37.4(0.98)	**6.72(0.332)**	70.5(0.26)	**65.4(0.152)**	58.7(0.313)
	GCE	73.1(0.272)	62.5(0.542)	33.7(0.599)	4.35(0.65)	**70.6(0.126)**	64.9(0.14)	58.2(0.249)
	TGCE	71.6(0.087)	58.8(0.229)	31.3(0.514)	4.84(0.498)	68.9(0.311)	62.8(0.174)	53.8(0.432)
	WW	62.6(0.248)	49.8(0.357)	31.7(0.284)	4.54(0.23)	59.8(0.446)	55.8(0.224)	44.6(0.548)
	JS	61.5(0.403)	37.1(0.774)	15.2(0.892)	5.51(0.464)	55.8(0.842)	40.4(0.241)	26.6(1.45)
	CS	3.98(0.328)	2.69(0.165)	2.5(0.0)	2.5(0.0)	3.27(0.303)	2.76(0.09)	2.7(0.242)
	RLL	15.2(1.48)	10.9(0.915)	8.34(0.455)	4.55(1.0)	12.3(0.401)	10.2(0.751)	7.71(0.358)
	NCE+RCE	6.78(0.139)	6.48(0.493)	5.22(0.178)	2.79(0.345)	6.43(0.506)	5.95(0.207)	5.0(0.217)
	NCE+AUL	7.72(0.263)	6.74(0.248)	5.65(0.641)	3.5(0.525)	7.41(0.272)	6.72(0.265)	5.76(0.346)
	NCE+AGCE	7.65(0.408)	7.01(0.278)	6.21(0.345)	3.78(0.382)	7.75(0.279)	6.79(0.318)	5.77(0.178)
	MSE	9.11(0.375)	4.59(0.356)	4.18(0.341)	3.28(0.109)	4.76(0.629)	4.55(0.27)	4.22(0.093)
	MAE	2.57(0.107)	3.97(0.31)	3.46(0.364)	3.31(0.478)	4.14(0.354)	4.02(0.241)	3.92(0.13)

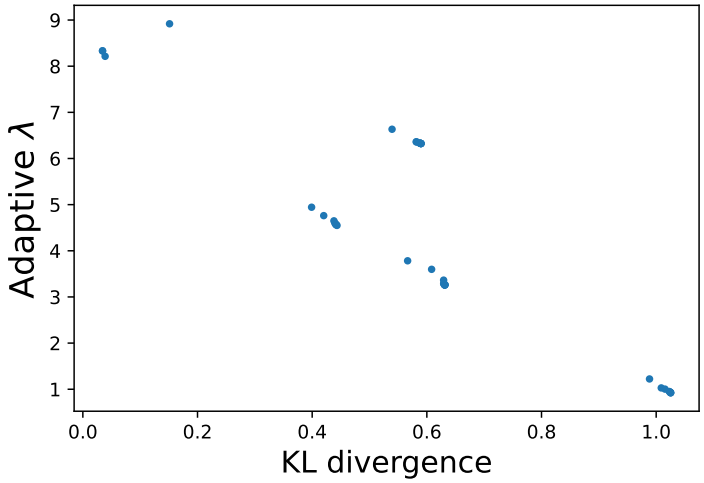


Figure 3. The averaged λ_t and $\mathrm{KL}(\mathbf{p}_t, \frac{\mathbf{1}}{K})$ values across all iterations for each instance on the synthetic dataset.

L.4. Leaderboard with Standard Deviation

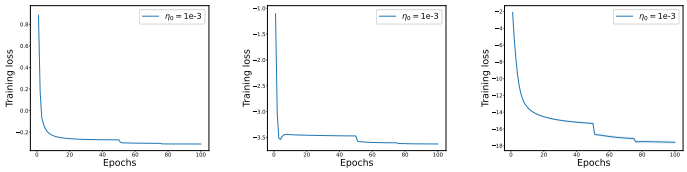
In addition to the averaged rank for the leaderboard at Table 2, we provide the leaderboard with standard deviation at Table 16.

Table 16. leaderboard for comparing 15 different loss functions on 7 datasets in 6 noise and 1 clean settings. The reported numbers are averaged ranks of performance. The smaller the better.

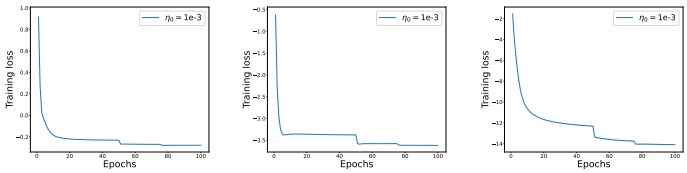
Loss Function	top-1	top-2	top-3	top-4	top-5	overall
ALDR-KL	**2.163(1.754)**	**2.469(2.149)**	**2.245(1.933)**	**2.224(1.951)**	**2.673(2.543)**	**2.355(2.092)**
LDR-KL	2.449(1.819)	2.571(1.841)	2.776(2.043)	3.0(2.138)	2.959(2.109)	2.751(2.006)
CE	4.714(1.852)	4.592(2.465)	4.224(2.131)	4.449(2.148)	4.388(2.184)	4.473(2.171)
SCE	4.959(1.958)	4.816(1.769)	4.449(1.63)	4.918(2.174)	4.469(1.63)	4.722(1.857)
GCE	5.796(2.39)	5.347(2.462)	6.061(2.18)	5.551(2.167)	5.571(2.241)	5.665(2.304)
TGCE	6.204(2.204)	6.204(1.948)	6.286(2.109)	6.204(1.84)	6.306(1.929)	6.241(2.011)
WW	7.673(2.198)	6.592(2.465)	6.061(2.377)	5.898(2.35)	5.469(2.492)	6.339(2.496)
JS	7.816(2.946)	7.837(2.637)	7.857(2.295)	7.98(2.114)	8.245(2.005)	7.947(2.43)
CS	8.653(4.943)	8.857(4.699)	8.755(4.506)	8.551(4.764)	8.714(4.84)	8.706(4.754)
RLL	10.184(2.569)	10.224(1.951)	10.224(2.41)	10.49(2.34)	10.408(2.166)	10.306(2.3)
NCE+RCE	9.694(3.065)	10.633(2.686)	10.878(2.553)	10.612(2.798)	10.714(2.483)	10.506(2.756)
NCE+AUL	10.449(2.167)	10.98(1.436)	11.102(1.297)	11.245(1.17)	11.163(1.633)	10.988(1.605)
NCE+AGCE	11.735(1.613)	11.837(1.448)	11.878(1.56)	11.51(1.704)	11.673(1.038)	11.727(1.496)
MSE	12.776(2.216)	12.612(2.64)	12.735(2.126)	12.939(1.921)	12.878(2.067)	12.788(2.21)
MAE	14.735(0.563)	14.429(1.666)	14.469(1.739)	14.429(2.06)	14.367(2.087)	14.486(1.72)

L.5. Convergence for ALDR-KL Optimization

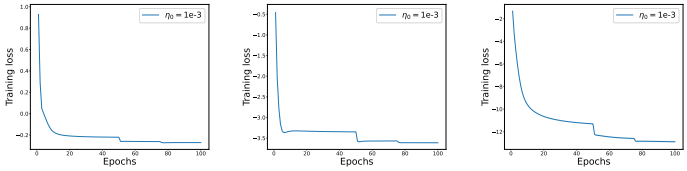
We show the practical convergence curves (mean value with standard deviation band) for Algorithm 1 for ALDR-KL loss at this section. 'CD' is short for class-dependent noise; 'U' is short for uniform noise; 'c' is the margin value; initial learning rate η_0 is tuned in $\{1e-1, 1e-2, 1e-3\}$. The convergence curves on ALOI dataset is shown in Figure 4; the convergence curves on News20 dataset is shown in Figure 5; the convergence curves on Letter dataset is shown in Figure 6; the convergence curves on Vowel dataset is shown in Figure 7.



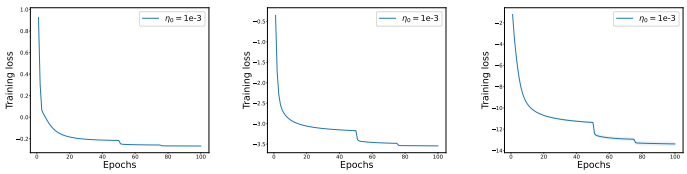
(a) CD(0.1), $c = 0.1$ (b) CD(0.1), $c = 1$ (c) CD(0.1), $c = 10$



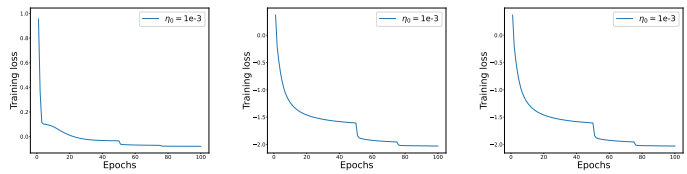
(d) CD(0.3), $c = 0.1$ (e) CD(0.3), $c = 1$ (f) CD(0.3), $c = 10$



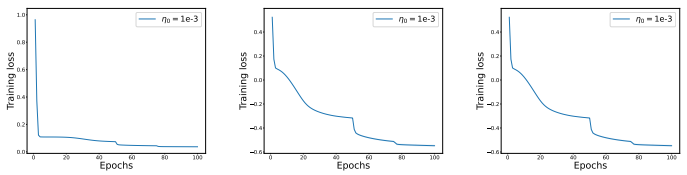
(g) CD(0.5), $c = 0.1$ (h) CD(0.5), $c = 1$ (i) CD(0.5), $c = 10$



(j) U(0.3), $c = 0.1$ (k) U(0.3), $c = 1$ (l) U(0.3), $c = 10$

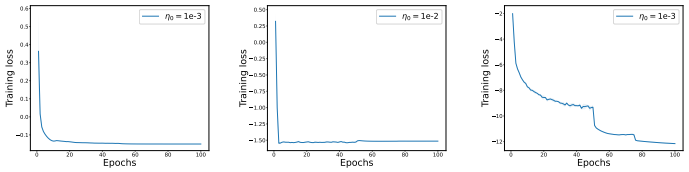


(m) U(0.6), $c = 0.1$ (n) U(0.6), $c = 1$ (o) U(0.6), $c = 10$

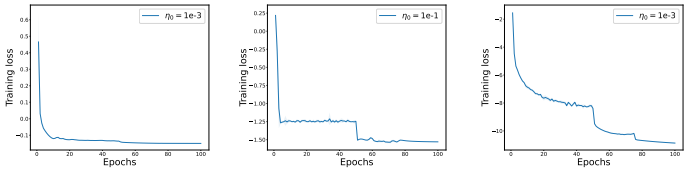


(p) U(0.9), $c = 0.1$ (q) U(0.9), $c = 1$ (r) U(0.9), $c = 10$

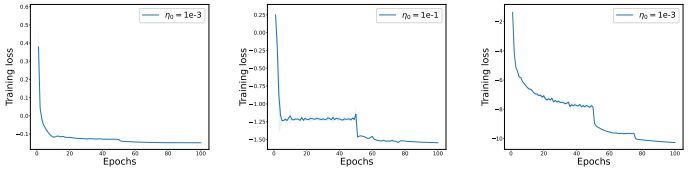
Figure 4. Training loss convergence for ALDR-KL (Algorithm 1) on ALOI dataset



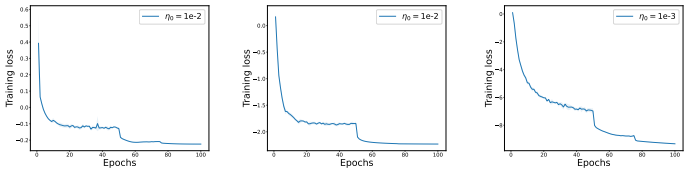
(a) CD(0.1), $c = 0.1$ (b) CD(0.1), $c = 1$ (c) CD(0.1), $c = 10$



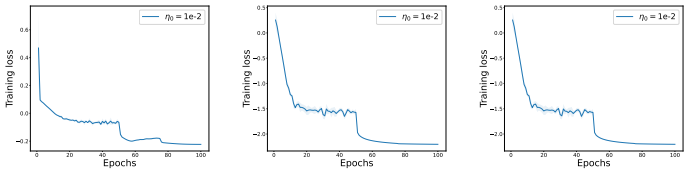
(d) CD(0.3), $c = 0.1$ (e) CD(0.3), $c = 1$ (f) CD(0.3), $c = 10$



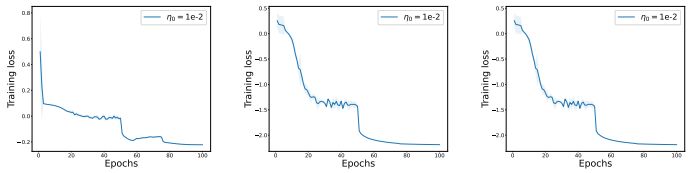
(g) CD(0.5), $c = 0.1$ (h) CD(0.5), $c = 1$ (i) CD(0.5), $c = 10$



(j) U(0.3), $c = 0.1$ (k) U(0.3), $c = 1$ (l) U(0.3), $c = 10$

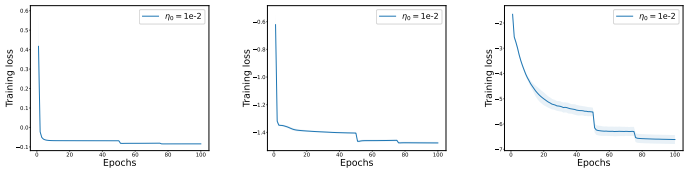


(m) U(0.6), $c = 0.1$ (n) U(0.6), $c = 1$ (o) U(0.6), $c = 10$

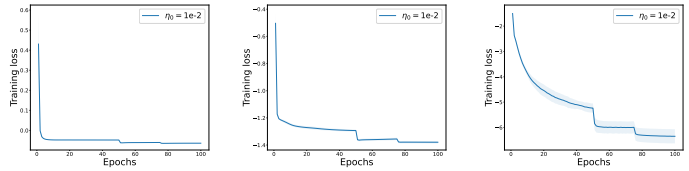


(p) U(0.9), $c = 0.1$ (q) U(0.9), $c = 1$ (r) U(0.9), $c = 10$

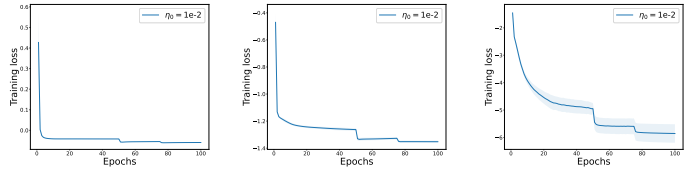
Figure 5. Training loss convergence for ALDR-KL (Algorithm 1) on News20 dataset



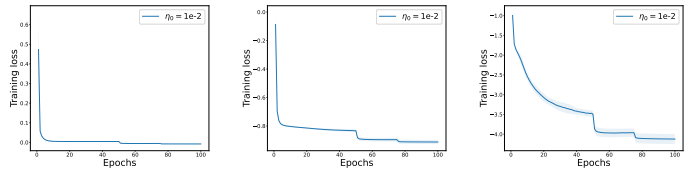
(a) CD(0.1), $c = 0.1$ (b) CD(0.1), $c = 1$ (c) CD(0.1), $c = 10$



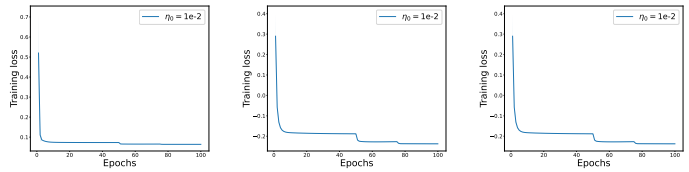
(d) CD(0.3), $c = 0.1$ (e) CD(0.3), $c = 1$ (f) CD(0.3), $c = 10$



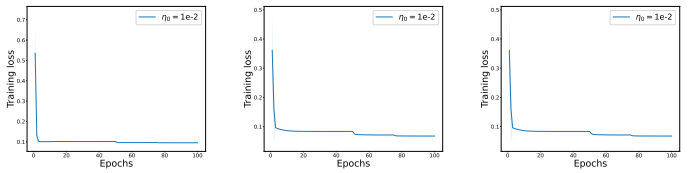
(g) CD(0.5), $c = 0.1$ (h) CD(0.5), $c = 1$ (i) CD(0.5), $c = 10$



(j) U(0.3), $c = 0.1$ (k) U(0.3), $c = 1$ (l) U(0.3), $c = 10$

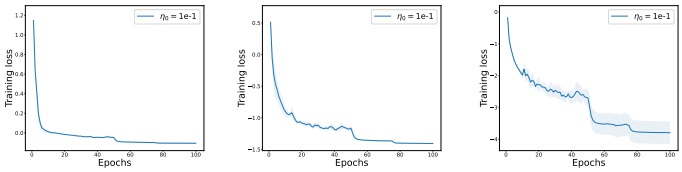


(m) U(0.6), $c = 0.1$ (n) U(0.6), $c = 1$ (o) U(0.6), $c = 10$

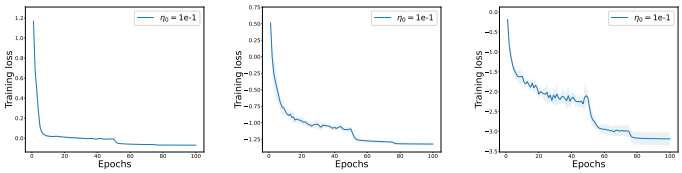


(p) U(0.9), $c = 0.1$ (q) U(0.9), $c = 1$ (r) U(0.9), $c = 10$

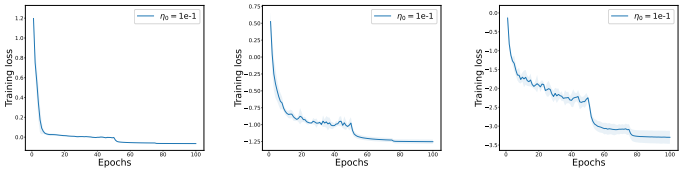
Figure 6. Training loss convergence for ALDR-KL (Algorithm 1) on Letter dataset



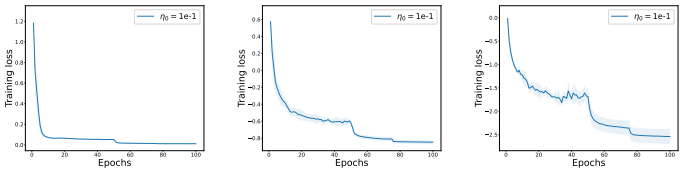
(a) CD(0.1), $c = 0.1$ (b) CD(0.1), $c = 1$ (c) CD(0.1), $c = 10$



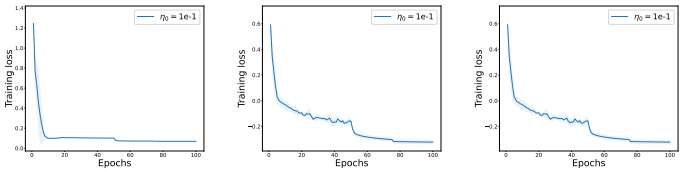
(d) CD(0.3), $c = 0.1$ (e) CD(0.3), $c = 1$ (f) CD(0.3), $c = 10$



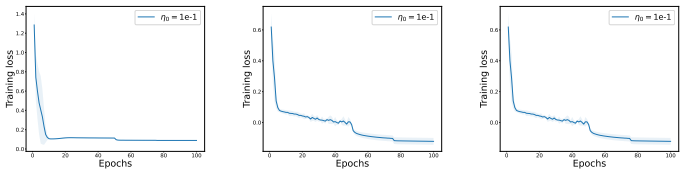
(g) CD(0.5), $c = 0.1$ (h) CD(0.5), $c = 1$ (i) CD(0.5), $c = 10$



(j) U(0.3), $c = 0.1$ (k) U(0.3), $c = 1$ (l) U(0.3), $c = 10$



(m) U(0.6), $c = 0.1$ (n) U(0.6), $c = 1$ (o) U(0.6), $c = 10$



(p) U(0.9), $c = 0.1$ (q) U(0.9), $c = 1$ (r) U(0.9), $c = 10$

Figure 7. Training loss convergence for ALDR-KL (Algorithm 1) on Vowel dataset

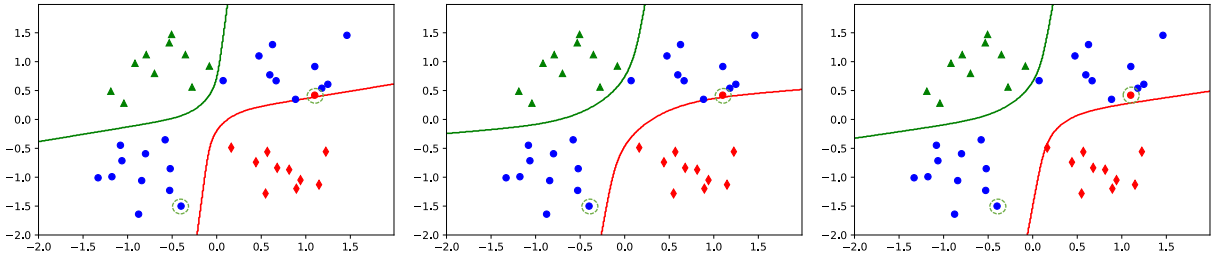


Figure 8. Design: training from scratch. Left: decision boundary for CE loss; Middle: decision boundary for LDR-KL loss with λ set as $\mathbb{E}_{\mathbf{x}_i,t}[\lambda_t^i]$ from ALDR-KL loss; Right: decision boundary for ALDR-KL loss.

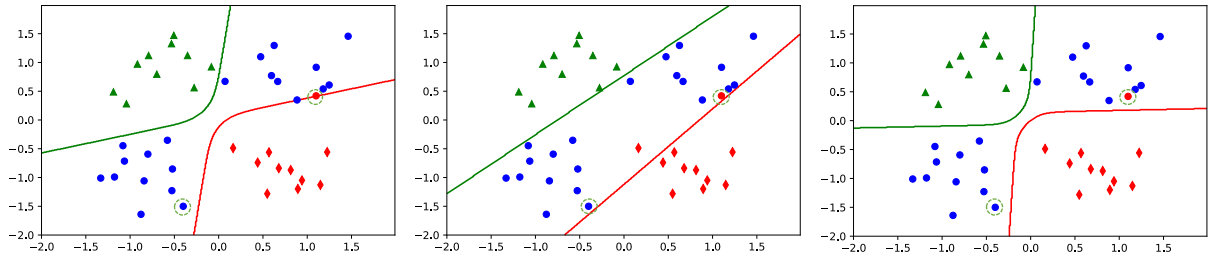


Figure 9. Design: pretrained from CE loss on clean data. Left: decision boundary for pretrained model with CE loss; Middle: decision boundary for LDR-KL loss with λ set as $\mathbb{E}_{\mathbf{x}_i,t}[\lambda_t^i]$ from ALDR-KL loss; Right: decision boundary for ALDR-KL loss.

L.6. More Synthetic Experiments

We additionally provide the synthetic experiments for training from scratch setting in Figure 8, and LDR-KL loss in Figure 9 and Figure 8. The experimental settings are kept as the same, except for training from scratch setting, the total training epoch is fixed as 200 for all methods. The λ parameter for LDR-KL loss is set as the mean value from the optimized λ from ALDR-KL loss. From the results, the ALDR-KL loss consistently enjoys the most robust decision boundary.

L.7. Running Time Cost

We provide the empirical running time cost for all baselines in Table 17. Each entry stands for mean and standard deviation for 100 consecutive epochs running on a x86_64 GNU/Linux cluster with NVIDIA GeForce GTX 1080 Ti GPU card.

Table 17. Running time for 15 baselines on 7 benchmark datasets (seconds per epoch).

Loss	Vowel	Letter	News20	ALOI	Kuzushiji-49	CIFAR100	Tiny-ImageNet
ALDR-KL	0.176(0.009)	0.933(0.033)	2.002(0.068)	7.029(0.091)	121.123(1.767)	31.923(0.381)	61.204(0.26)
LDR-KL	0.172(0.006)	0.783(0.033)	1.868(0.107)	5.779(0.098)	116.774(0.512)	31.507(0.826)	59.106(0.318)
CE	0.17(0.008)	0.719(0.027)	1.846(0.095)	5.157(0.074)	116.612(0.296)	31.49(0.289)	58.687(0.141)
SCE	0.167(0.008)	0.756(0.024)	2.082(0.081)	5.903(0.175)	120.376(0.275)	29.97(0.156)	60.336(0.456)
GCE	0.165(0.008)	0.771(0.029)	1.869(0.079)	5.772(0.42)	122.294(0.268)	30.373(0.187)	61.026(0.213)
TGCE	0.178(0.009)	0.769(0.03)	1.921(0.077)	5.813(0.193)	120.343(1.339)	30.293(0.163)	58.869(0.765)
WW	0.155(0.009)	0.715(0.028)	2.067(0.068)	5.255(0.165)	127.304(0.49)	29.386(0.148)	63.249(0.666)
JS	0.161(0.011)	0.876(0.034)	1.942(0.108)	6.648(0.1)	118.172(0.279)	31.618(0.259)	59.562(0.14)
CS	0.158(0.012)	0.73(0.036)	1.909(0.084)	5.393(0.178)	120.702(1.923)	29.278(0.139)	60.391(1.28)
RLL	0.172(0.008)	0.794(0.024)	2.118(0.077)	6.198(0.157)	122.021(0.339)	30.35(0.48)	60.609(0.785)
NCE+RCE	0.177(0.007)	0.818(0.037)	2.264(0.085)	6.326(0.194)	116.682(1.909)	30.139(0.124)	59.967(0.589)
NCE+AUL	0.174(0.007)	0.835(0.031)	2.154(0.121)	6.464(0.194)	115.613(0.766)	29.951(0.12)	60.022(0.49)
NCE+AGCE	0.169(0.01)	0.835(0.039)	2.21(0.122)	6.418(0.203)	121.702(1.55)	29.632(0.063)	63.533(0.499)
MSE	0.158(0.006)	0.702(0.025)	1.932(0.089)	4.983(0.158)	117.601(1.364)	29.951(0.46)	59.44(0.684)
MAE	0.163(0.011)	0.687(0.031)	1.884(0.263)	4.843(0.179)	119.065(1.854)	29.404(0.092)	61.176(1.256)

L.8. Ablation Study for λ and λ_0

We addtionally show the validation results for LDR-KL and ALDR-KL losses on CIFAR100 dataset with different λ and λ_0 values in Table 18. When noisy level is higher, LDR-KL or ALDR-KL loss prefer larger λ or λ_0 (1 over 0.1) on CIFAR100 dataset.

Table 18. The mean and standard deviation for top-1 validation accurcay on CIFAR100 dataset. The best choice is marked as bold.

Loss	LDR-KL			ALDR-KL		
Noise setting	λ=0.1	λ=1	λ=10	λ_0=0.1	λ_0=1	λ_0=10
Clean	**68.948(0.211)**	66.848(0.465)	61.528(0.665)	**68.456(0.351)**	66.54(0.304)	60.702(0.332)
Uniform 0.3	**40.658(0.218)**	38.918(0.337)	36.914(0.466)	**40.872(0.259)**	38.888(0.24)	36.958(0.255)
Uniform 0.6	13.462(0.831)	**16.084(0.467)**	14.766(0.306)	13.196(0.821)	**16.038(0.55)**	14.976(0.158)
Uniform 0.9	1.414(0.086)	**1.46(0.088)**	1.348(0.063)	1.444(0.084)	**1.468(0.044)**	1.366(0.118)
CD 0.1	**59.346(0.258)**	56.478(0.376)	51.696(0.718)	**59.368(0.32)**	55.99(0.328)	52.252(0.416)
CD 0.3	**43.198(0.434)**	40.098(0.375)	33.89(1.045)	**43.272(0.34)**	39.66(0.445)	34.354(1.212)
CD 0.5	32.026(0.373)	**32.344(0.373)**	25.672(0.379)	31.916(0.286)	**32.378(0.333)**	26.396(0.637)