A Unified DRO View of Multi-class Loss Functions with top-N Consistency

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-class classification is one of the most common tasks in machine learning applications, where data is labeled by one of many class labels. Many loss functions have been proposed for multi-class classification including two well-known ones, namely the cross-entropy (CE) loss and the crammer-singer (CS) loss (aka. the SVM loss). While CS loss has been used widely for traditional machine learning tasks for structured data, CE loss is usually a better choice (the default choice) for multi-class deep learning tasks. There are also top-k variants of CS loss and CE loss that are proposed to promote the learning of a classifier for achieving better top-k accuracy. Nevertheless, it still remains unclear the relationship between these different losses, which hinders our understanding of their expectations in different scenarios. In this paper, we present a unified view of the CS/CE losses and their smoothed top-k variants by proposing a new family of loss functions, which are arguably better than the CS/CE losses when the given label information is incomplete and noisy. The new family of smooth loss functions named label-distributionally robust (LDR) loss is defined by leveraging the distributionally robust optimization (DRO) framework to model the uncertainty in the given label information, where the uncertainty over true class labels is captured by using distributional weights for each label regularized by a function. We have two observations: (i) the CS and the CE loss are just two special cases of the LDR loss by choosing two particular values for the involved regularization parameter; hence the new LDR loss provides an interpolation between the CS loss and the CE loss, and also induces new variants; (ii) the smoothed top-k losses are also special cases of the LDR loss by regularizing the involved uncertainty variables into a bounded ball. Theoretically, we establish the top-N consistency (for any $N \geq 1$) of the proposed LDR loss, which is not only consistent with existing consistenty results for the CS and the CE loss but also addresses some open problems regarding the consistency of top-k SVM losses. Empirically, we provide some experimental results on synthetic data and real-world benchmark data to validate the effectiveness of the new variants of LDR loss.

1 Introduction

Multi-class classification is one of the fundamental tasks in machine learning (ML), where data is labeled by one of many class labels. In the past decades, multi-class classification has been used in a variety of areas, including image recognition [1, 2, 3], handwritten recognition [4, 5], natural language processing [6], etc. More training data could always be a possible solution to improve the model performance. However, labeling efforts could be more expensive as the machine learning task becomes more complicated. Furthermore, for multi-class classification, the labels of data could form a hierarchical tree structure, which leads to multi-label or ambiguity issue. As a consequence, for

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

computer vision task or natural language processing task, one image or text could contain multiple objects or concepts, which raises challenges not only for human labeling jobs but also model learning consistency [7, 8, 9]. On the other hand, a loss function plays a central role that can be researched for learning a model to achieve the learning consistency. As we cannot directly optimize the non-continuous 0-1 loss, optizming some surrogate loss functions such as the cross entropy (CE) and the SVM loss becomes the main-stream machine learning solution for multi-class classification task.

In this paper, we consider a machine learning solution based on optimizing a novel loss function and its top-k simplex constrained variant for improving the performance of the standard CE loss and the SVM loss. The motivation of our work springs from an interesting observation and its explanation: we have seen a lot of studies in traditional ML literature using the SVM losses (e.g., the Crammer-Singer (CS) loss [10]), but almost all studies of deep learning (DL) on ImageNet uses the CE loss. What is the underlying reason for this? One could explain this from two perspectives: (i) the CE loss is smooth while the CS loss is non-smooth, which is not favorable for optimization; (ii) 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 in an exceptional condition that the maximum conditional probability of a class label given the input is larger than 0.5 [11, 12]. The exceptional condition means that the data is associated with a single label with high probability. However, such condition usually does not hold for real-world multi-class classification tasks in DL applications where data in inherently associated with multiple class labels. Although each image in ImageNet is labeled by only one class, the recent studies [7, 8] show that these images inherently belong to multiple class labels. In particular, they observed that a significant fraction of images—more than a fifth—contains at least two objects from the 1000 classes.

2 Related Work

Multi-class Loss functions. Multi-class loss functions have been studied extensively in conventional machine learning (ML) literature [13, 14, 11, 10]. Many loss functions are shown to be consistent (producing Bayes optimal classifier for minimizing zero-one loss, i.e., top-1 error) (e.g., CE loss), while others are shown to be inconsistent (e.g., CS loss, WW loss). Variants of these standard losses have been developed for minimizing top-k error ($k > 1$), such as (smoothed) top-k SVM loss, top-k entropy loss [15, 16, 17, 18]. However, the relationship between these losses and the CE loss is still unclear, which hinders the understanding of their performance in different scenarios. 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 bring new loss functions. In particular, the proposed LDR-KL loss has arguably more advantages than the CE loss and our methods based on the LDR-KL loss yields consistently better performance than the CE loss on benchmark datasets.

Distributional robust optimization (DRO) Objective. It is interesting to connect the proposed LDR loss to the DRO objective in previous works for handling noisy and imbalanced data [19, 20, 21, 22]. Different from the proposed LDR loss that assocaite the DW variables to class labels, the existing DRO objective assigns DW variables to instances and defines an aggregate objective by maximizing over the DW variables in an uncertainty set. But it is interesting to make the analogy between the LDR loss and the DRO objective for arguing the benefit of the LDR loss on handling the incomplete and noisy label information (cf. the discussion in section Label-Distributionally Robust (LDR) Losses).

3 The Proposed Method

Notations. Let $\mathbf{x} \in \mathbb{R}^d$ denote an input data, $y \in [K] := \{1, \ldots, K\}$ denote its class label, and let $f(\mathbf{x}) \in \mathbb{R}^K$ denote the prediction scores of a model. Without causing any confusion, we simply use the notation $\mathbf{f} = (f_1, \ldots, f_K) = f(\mathbf{x}) \in \mathbb{R}^K$ 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.

3.1 Label-Distributionally Robust (LDR) Losses

The motivation of the proposed LDR losses is that the given label y for a given data $\mathbf{x}$ is usually incomplete. Hence, the label information encoded in $\mathbf{y}$ that is used to define a loss function is

noisy, meaning that the zero entries in $\mathbf{y}$ are not necessarily true zero. In addition, the given label y itself could be wrong due to the labeling error, meaning that the entry in $\mathbf{y}$ that is equal to 1 is not necessarily true one. 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. 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:

$$\psi_\lambda(\mathbf{f}, y) = \max_{\mathbf{p}\in\Omega} \sum_{k=1}^{K} p_k(f_k - f_y + c(y,k)) - \lambda R(\mathbf{p}) \tag{1}$$

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 $\mathbf{p}$, $c(y,k) = c_y \mathbb{I}(y \neq k)$ denotes the margin parameter, $\lambda > 0$ is the DW regularization parameter, and $R(\mathbf{p})$ is is a regularization term of $\mathbf{p}$ which is set by a **strongly convex function**. The strong convexity of $R(\mathbf{p})$ makes $\psi_\lambda(\mathbf{f}, y)$ a smooth function of $\mathbf{f}$ due to the duality between strong convexity and smoothness [23].

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., $R(\mathbf{p}) = \sum_{k=1}^{K} p_i \log(p_i K)$. In this case, we can derive a closed-form expression of $\psi_\lambda(\mathbf{f}, y)$, denoted by $\psi_\lambda^{\text{KL}}(\mathbf{f}, y)$,

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

It is notable that the DW regularization parameter λ is a key feature of this loss. It is interesting to see that the LDR-KL loss covers several existing losses as special case or extreme cases by varying the value of λ. We include the proof in the appendix.

Extreme Case I. $\lambda = 0$. The loss function becomes: $\psi_\lambda^{\text{CS}}(\mathbf{f}, y) = \max(0, \max_{k\neq y} f_k + c_y - f_y)$ which is known as Crammer-Singer loss (an extension of hinge loss) for multi-class SVM [10] .

Extreme Case II. $\lambda = \infty$. The loss function becomes: $\psi_\lambda^{+}(\mathbf{f}, y) = \frac{1}{K}\sum_{k=1}^{K}(f_k - f_y + c(y,k))$. This loss is similar to the Weston-Watkins (WW) loss [13] for multi-class SVM, which is given by $\sum_{k\neq y} \max(0, f_k - f_y + 1)$.

Special Case. With $\lambda = 1$, the LDR-KL loss becomes the CE loss: $\psi_\lambda^{\text{CE}}(\mathbf{f}, y) = \log\left(\sum_k \exp\left(f_k + c(y,k) - f_y\right)\right)$.

Note that in standard CE loss, the margin parameter c_y is 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 [24] for imbalanced data classification with different margin values c_y for different classes.

3.2 Top-N Consistency

The LDR-KL loss: the top-N error is defined as $\text{err}_N(\mathbf{f}, y) = \mathbf{I}(y \notin r_N(\mathbf{f}))$, where r_N is a top-N selector that selects the N indices of the largest entries of the input by breaking ties arbitrarily. According to previous studies [18], a necessary condition for a multi-class loss function is top-N consistent is that the loss function is top-N calibrated define below. Hence, we will only present the result of top-N calibration of the LDR-KL loss, and its implication on consistency can be found in previous studies (cf. [18]). Let $P_N(\mathbf{f}, \mathbf{q})$ denotes that $\mathbf{f}$ is top-N preserving with respect to $\mathbf{q}$, i.e., if for all $l \in [K]$, $q_l > q_{[N+1]} \Longrightarrow f_l > f_{[N+1]}$, and $q_l < q_{[N]} \Longrightarrow f_l < f_{[N]}$. For anys $\mathbf{q} \in \Delta_K$, let $L_\psi(\mathbf{f}, \mathbf{q}) = \mathrm{E}_{y\sim\mathbf{q}}\psi_y(\mathbf{f}) = \sum_l q_l \psi(\mathbf{f}, l)$.

Definition 1. *[18] A loss function $\psi(\mathbf{f}, y) : \mathbb{R}^K \times [K] \to \mathbb{R}$ is called top-N calibrated for all $\mathbf{q} \in \Delta_K$ it holds that:* $\inf_{\mathbf{f}\in\mathbb{R}^K : \neg P_N(\mathbf{f},\mathbf{q})} L_\psi(\mathbf{f}, \mathbf{q}) > \inf_{\mathbf{f}\in\mathbb{R}^K} L_\psi(\mathbf{f}, \mathbf{q})$.

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 consistent with respect to $\mathbf{q}$, i.e., satisfying that for any pair $q_i < q_j$ it holds that $\mathbf{f}_i^* < \mathbf{f}_j^*$, then $\psi_y(\mathbf{f})$ is top-N calibrated for any $N \geq 1$.*

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 consistent for any $\lambda > 0$ and hence the LDR-KL loss $\psi_\lambda^{KL}(\mathbf{f}, y)$ with any $\lambda > 0$ is top-N calibrated for any $N \geq 1$.*

We include the proof of the above lemma and theorem in the appendix.

Consistency of the general LDR losses: we present a sufficient condition on the DW regularization function R such that the resulting LDR loss $\psi_\lambda(\mathbf{f}, y)$ is consistent. The result derived below will address an open problem in [17] regarding the consistency of smoothed top-k SVM loss.

Definition 2. *A function $R(\mathbf{p})$ is 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 symmetric, R is strongly convex satisfying $\partial R(0) < 0$ and $[\partial R(p)]_i > 0 \Longrightarrow p_i \neq 0$, then $\psi_\lambda(\mathbf{f}, y)$ is top-N classification calibrated for any $N \geq 1$.*

Consistency of LDR-k loss. As an application of our result above, we consider the LDR-k loss for a given value $k \in [K]$. In particular, by specifying top-k simplex $\Omega = \Omega(k) = \{\mathbf{p} \in \mathbb{R}^K : \sum_{l=1}^K p_l \leq 1, p_l \leq 1/k, \forall l\}$ and $\lambda = 0$, the LDR-k loss becomes the top-k SVM loss [17], i.e.,

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

where $\mathbf{c}_y = (c(y, 1), \ldots, c(y, K))$. This top-$k$ SVM loss is not consistent [17]. But, we can make it consistent by adding a strongly convex regularizer $R(\mathbf{p})$ satisfying the conditions in Theorem 2. In particular, we define a **LDR-k loss** by

$$\psi_\lambda^k(\mathbf{f}, y) = \max_{\mathbf{p}\in\Omega(k)} \sum_l p_l(f_l - f_y + c(y, l)) - \lambda R(\mathbf{p}) \tag{2}$$

Theorem 2 implies $\psi_y^k(\mathbf{f})$ is classification calibrated and hence consistent. This addresses the open problem raised in [17] regarding the consistency of smoothed top-k SVM loss for multi-class classification.

3.3 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 (2), 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. Due to the limited space, we include the details of the **Theorem 3** and the proof in the appendix.

4 Experiments

In this section, we first show the effectiveness of the proposed LDR approach on a synthetic dataset and then verify the effectiveness of our proposed LDR and LDR-k methods on some well studied benchmark datasets for multiclass classification tasks. We also conduct experiments on image datasets using pretrained deep learning based features for learning a linear model with details described shortly. We compare our LDR-KL and LDR-k-KL losses with the CE loss (with margin parameters), the CS loss, the WW loss, the top-k focused hinge loss (SVM-k) [18] $\psi^{\text{svm}-k}(\mathbf{f}, y) = (f_{\pi(k)} - f_y + c(y, \pi(k)))_+$, the top-$k$ truncated CE loss (CE-k) [25] $\psi^{\text{CE}-k}(\mathbf{f}, y) = \log\left(1 + \sum_{j=k}^{K-1} \exp(f_{[j]}^{\setminus y} - f_y)\right)$. Among of them, CE, SVM-k, LDR-KL and LDR-k-KL are calibrated for top-k consistency. The details are included in the appendix.

5 Conclusion and Future Work

In this paper, we have proposed a novel family of label distributional robust losses that covers the cross entropy and the SVM loss as special cases and also induces new smoothed losses. We proved the consistency of the proposed losses. Our empirical studies also demonstrate the usefulness of the proposed loss functions. For future work, we will consider to learn individual λ for each example in order to capture the difference between different examples in the given label information.

References

[1] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 and CIFAR-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6:1, 2009.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[3] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

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

[5] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[6] Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier, 1995.

[7] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *CoRR*, abs/2006.07159, 2020.

[8] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *ArXiv preprint arXiv:2005.11295*, 2020.

[9] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5, 2004.

[10] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, March 2002.

[11] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*, 5:1225–1251, December 2004.

[12] Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *Transactions on Pattern Analysis and Machine Intelligence*, 2018.

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

[14] Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(May):1007–1025, 2007.

[15] Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass svm. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 325–333. Curran Associates, Inc., 2015.

[16] Maksim Lapin, Matthias Hein, and Bernt Schiele. 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*, pages 1468–1477. IEEE Computer Society, 2016.

[17] Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(7):1533–1554, 2018.

[18] Forest Yang and Sanmi Koyejo. On the consistency of top-k surrogate losses. In *International Conference on Machine Learning*, pages 10727–10735. PMLR, 2020.

[19] Shai Shalev-Shwartz and Yonatan Wexler. Minimizing the maximal loss: How and why. In *ICML*, pages 793–801, 2016.

[20] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pages 2971–2980, 2017.

[21] Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. A practical online method for distributionally deep robust optimization. *arXiv preprint arXiv:2006.10138*, 2020.

[22] Dixian Zhu, Zhe Li, Xiaoyu Wang, Boqing Gong, and Tianbao Yang. A robust zero-sum game framework for pool-based active learning. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 517–526, 2019.

[23] Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: learning applications and matrix regularization. Technical report, Toyota Technological Institute, 2009.

[24] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 1567–1578, 2019.

[25] M. Lapin, M. Hein, and B. Schiele. Loss functions for top-k error: Analysis and insights. In *CVPR*, pages 1468–1477, 2016.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

A Proof for The LDR-KL Loss Formulation

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

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

It is equivalent to:

$$\lambda\big(\max_{\mathbf{p}\in\Delta} \frac{\mathbf{q}}{\lambda} - \sum_i \mathbf{p}_i \log K\mathbf{p}_i\big)$$

We are interested in the component inside of the parenthesis. Apply Lagrangian multiplier:

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

Stationary condition of $\mathbf{p}$:

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

Dual form:

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

Stationary condition for β :

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

Plug in β^* back to dual form:

$$1 + \log \sum_i \frac{1}{K} \exp(\frac{\mathbf{q}_i}{\lambda} - 1)$$

□

B Proof for Theorem 1

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

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

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

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

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

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

Moreover:

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

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

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

If $c_k = c$, we have

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

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 $f_i^* < f_j^*$. □

C Proof of Theorem 2

Definition 3. *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 symmetric, then R^* is also 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

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

Then

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

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

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

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 $\mathbf{f}^\sharp \in \partial R(\mathbf{p}^*)$. Since $[\mathbf{f}^\sharp]_2 > 0$, we have $p_2^* > 0$. 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 $R^*(\mathbf{f}^\sharp) - R^*(\mathbf{f}^\ddagger) < 0$.

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

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

Hence

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

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 $[\mathbf{f} - f_l\mathbf{1} - c_l]_1 = [\mathbf{f} - f_l\mathbf{1} - c_l]_2$. As a result,

$$q_1 \mathbf{1}^\top \nabla R^*(\mathbf{f} - f_1\mathbf{1} - c_1) = q_2 \mathbf{1}^\top \nabla R^*(\mathbf{f} - f_2\mathbf{1} - c_2)$$

It is clear that the above equation is impossible since $\mathbf{1}^\top \nabla R^*(\mathbf{f}) \neq 0, \forall \mathbf{f}$ with $[f]_1 = 0$ and $q_1 < q_2$. □

D Theorem 3

Theorem 3. *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* $\pi(i)$ *denote the index for its* i*-th largest value. Given* $a \in [K]$ *define a vector* $\mathbf{p}(a)$ *as*

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

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

Given that the algorithm would be simple with the presented theorem 3, we do not present a formal algorithm box here. Instead, we simply describe it as follows: i) scan (from largest to smallest) through vector $\mathbf{q}$ to get cumulative summation $\sum_{j=a}^{K} \exp(\frac{\mathbf{q}_{\pi(j)}}{\lambda} - 1)$ for all possible values of a. ii) apply the theorem 3 to calculate $[\mathbf{p}(a)]_{\pi(a)}$ 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 $\pi(a)$ that $\boldsymbol{\gamma}^*_{\pi(a)} = 0$ and $\boldsymbol{\gamma}^*_{\pi(a-1)} > 0$, where $2 \leq a \leq k+1$, (or $a = 1$, $\boldsymbol{\gamma}^*_{\pi(a-1)}$ is undefined):

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

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

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

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

For those $i < a$ that $\gamma^*_{\pi(i)} > 0$:

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

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

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

- Suppose $a' < a$: because $\forall a' \ s.t. \ k+1 \geq a > a' \geq 1$, assume $[\mathbf{p}(a')]_{\pi(a')} \leq \frac{1}{k} \implies a'$ is another smaller valid a, which violates the pre-condition; therefore, $[\mathbf{p}(a')]_{\pi(a')} > [\mathbf{p}(a)]_{\pi(a')} = \frac{1}{k}$, 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 $[\boldsymbol{\gamma}(a)]^*_{\pi(a)} = 0$ and $[\boldsymbol{\gamma}(a')]^*_{\pi(a)} > 0$, 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 $[\boldsymbol{\gamma}(a)]^*_{\pi(i)} = [\boldsymbol{\gamma}(a')]^*_{\pi(i)} > 0, \forall i < a$ in order to hold the $\Omega(k)$ constrain).

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

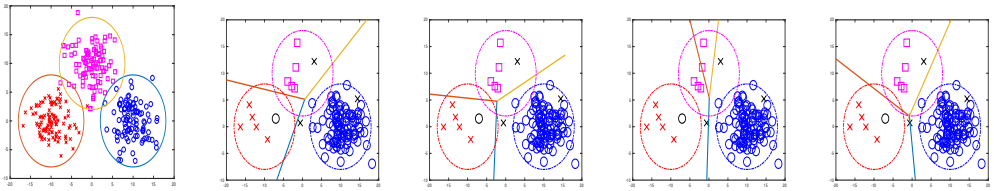
  On the other hand, $\forall j < a', \ [\mathbf{p}(a')]_{\pi(j)} = \frac{1}{k} \geq [\mathbf{p}(a)]_{\pi(j)} \implies \sum_{i=a'}^{K}[\mathbf{p}(a')]_{\pi(i)} \leq 1 - \sum_{i=1}^{a'-1}[\mathbf{p}(a')]_{\pi(i)} \leq 1 - \sum_{i=1}^{a'-1}[\mathbf{p}(a)]_{\pi(i)} = \sum_{i=a'}^{K}[\mathbf{p}(a)]_{\pi(i)}$,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

□

E Experiments Details

E.1 Synthetic Data

For multi-class classification, all the loss functions are defined in a sense to predict the data closer to the labeled class than other classes. However, one interesting but rarely considered question would be: how uncertain could another unlabeled class possibly be a potential ground truth for the data? On the one hand, CS loss is defined in a way that all other unlabeled classes should have equally chance to be a potential true class because only the most wrongly classified class is penalized by CS loss. On the other hand, WW loss is defined in a way that all other unlabeled classes are penalized equally; therefore, the model is spun up to infer the potential unlabeled true class by utilizing all the labeled training data. LDR loss takes a trade off between CS and WW loss from the potential unlabeled true classes discovering perspective because the DW regularization controls whether LDR loss is closer to



(a) True Data Dis. (b) w/ CE loss (c) w/ CS loss (d) w/ WW loss (e) w/ LDR loss

Figure 1: The CE loss is vulnerable to the imbalanced classes case, as we can see the major class (blue circle) occupying the majority of the space on the plane, which disagrees the boundaries of underlying distributions. The CS loss is also sensitive to the imbalanced classes in a sense that the intersection of the hyper-planes is pushed away from the major class. On the other hand, WW loss is sensitive to outliers, where the boundary between '$\times$' and '$\square$' is highly misled and closer to vertical direction because WW loss wants to turn up the probability that the outliers are classified to the given label against the other classes equally. LDR loss, in contrast, provides more reasonable boundaries in this example, which benefits from the uncertainty capturing by the distributional weight (DW) regularization.

Table 1: statistics for the benchmark datasets (imagenet does not have fixed features number; we manually crop/resize the images to make them consistent.)

Dataset	# of samples	# of features	# of classes
aloi	108000	128	1,000
news20	15935	62060	20
letter	15000	16	26
cifar10	60000	32x32x3	10
cifar100	60000	32x32x3	100
imagenet	1331167	224x224x3	1000

CS loss or WW loss. CE loss is a specific form of LDR loss. With DW regularization, LDR loss is closer to CE loss with adjustable temperature parameter. DW regularization also could be treated as a uncertainty capturing module for potential unlabeled true labels.

We justify the effectiveness of the proposed DW regularization for LDR loss on a synthetic data shown in Figure 1. We first define 3 bi-variate normal distributions, whose centers are $(10, 0)^\top$, $(-10, 0)^\top$, $(0, 10)^\top$ and co-variances are $\begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$. They are used to mimic the underlying distributions for 3 different classes. For training, we first sample 100 data samples for one class and 5 data samples for the other two classes, which is used to simulate imbalanced classes scenario. Next, we manually impose 4 outliers, which are either denoted as black '$\times$' and '$\circ$' as an analogy to similar images mislabeled as different classes for image classification task. They are either sitting on the different classes comparing to their underlying, or on the middle of two classes. For the purpose of justification for DW regularization, we only compare CE, CS, WW and LDR in this subsection and leave out their top-k variants. The DW regularization is tuned in the range {1e-1,1,1e1}, and ℓ_2 regularization is tuned in {1e-4,1e-3,1e-2,0} for all the methods. Margin is fixed as 1 for all the methods. The final model are selected by 5-fold cross-validation. The classification boundaries are drawn in Figure 1. As we can observe from the results, DW regularization improves robustness for LDR loss.

E.2 Benchmark Datasets

In this section, we conduct experiments on some benchmark datasets for multi-class classification. The statistics for the datasets are summarized in Table 1. The evaluation metrics include top-k accuracy ($k = \{1, ..., 5\}$).

Structured Data: We first compare different losses on the three structured datastesm, namely aloi, news20 and letter for learning a linear model [3, 6, 4][1]. We split 10% data samples for testing, the remaining for training and validation. It is worth noting that we utilize PCA for news20 dataset to keep 128 principle components, to avoid the high dimensionality issue.

[1]Data is available here: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

Table 2: Top Accuracy on six datasets with different loss functions

Data	Loss	top-1	top-2	top-3	top-4	top-5
aloi	CE	0.9052(2.3e-4)	0.9449(3.1e-4)	0.9586(2.0e-4)	0.9650(1.5e-4)	0.9711(1.5e-4)
	CS	0.9031(6.7e-4)	0.9450(7.3e-4)	0.9584(4.1e-4)	0.9661(3.4e-4)	0.9709(5.3e-4)
	WW	0.8355(1.4e-3)	0.9136(4.5e-4)	0.9394(8.2e-4)	0.9544(3.2e-4)	0.9634(6.6e-4)
	SVM-k	0.9031(6.7e-4)	0.9455(7.5e-4)	0.9610(8.7e-4)	0.9690(6.7e-4)	0.9742(6.1e-4)
	CE-k	0.9052(2.3e-4)	0.9412(4.0e-4)	0.9554(1.9e-4)	0.9632(1.8e-4)	0.9690(2.1e-4)
	LDR-KL	**0.9099(6.8e-4)**	0.9466(6.1e-4)	0.9595(4.1e-4)	0.9669(2.4e-4)	0.9714(1.9e-4)
	LDR-k-KL	**0.9099(6.8e-4)**	**0.9489(5.1e-4)**	**0.9664(4.0e-4)**	**0.9747(2.4e-4)**	**0.9784(2.4e-4)**
news20	CE	0.6353(1.7e-3)	0.7909(1.6e-3)	0.8633(7.3e-4)	0.9000(1.8e-3)	0.9331(5.3e-4)
	CS	**0.6579(1.5e-3)**	0.7960(1.3e-3)	0.8569(1.8e-4)	0.8905(8.8e-4)	0.9122(3.2e-4)
	WW	0.6403(3.1e-3)	0.7980(8.4e-4)	0.8663(1.5e-3)	0.9091(1.6e-3)	**0.9358(6.5e-4)**
	SVM-k	**0.6579(1.5e-3)**	0.7957(1.9e-3)	0.8673(4.0e-3)	**0.9097(3.6e-3)**	0.9321(1.8e-3)
	CE-k	0.6353(1.7e-3)	0.7942(3.2e-3)	0.8631(2.9e-3)	0.8948(1.3e-3)	0.9213(4.0e-3)
	LDR-KL	0.6501(1.3e-4)	0.7983(2.0e-3)	0.8632(1.3e-3)	0.9004(1.3e-3)	0.9284(8.7e-4)
	LDR-k-KL	0.6501(1.3e-4)	**0.8004(5.2e-4)**	**0.8680(0)**	0.9080(7.0e-4)	0.9328(4.9e-4)
letter	CE	0.7459(8.8e-4)	0.8485(2.7e-4)	0.8873(0)	0.9100(0)	0.9269(9.0e-4)
	CS	0.7589(3.3e-4)	0.8487(7.8e-4)	0.8812(2.7e-4)	0.8987(2.7e-4)	0.9153(0)
	WW	0.7141(1.4e-3)	0.8429(3.3e-4)	0.8920(0)	0.9182(3.3e-4)	0.9418(2.7e-4)
	SVM-k	0.7589(3.3e-4)	**0.8669(1.1e-3)**	**0.9073(1.4e-4)**	**0.9326(1.4e-3)**	**0.9649(2.9e-3)**
	CE-k	0.7459(8.8e-4)	0.8586(7.3e-4)	0.8913(1.5e-3)	0.9191(6.5e-4)	0.9423(1.2e-3)
	LDR-KL	**0.7603(2.7e-4)**	0.8580(0)	0.8887(0)	0.9073(0)	0.9260(0)
	LDR-k-KL	**0.7603(2.7e-4)**	0.8581(2.7e-4)	0.9008(4.2e-4)	0.9265(4.9e-4)	0.9453(5.9e-4)
CIFAR-10	CE	0.8950(1.2e-4)	0.9644(1.0e-4)	0.9847(0)	0.9915(4.0e-5)	0.9965(7.5e-5)
	CS	0.8944(1.4e-4)	0.9635(1.7e-4)	0.9845(1.5e-4)	0.9919(4.4e-5)	0.9965(4.9e-5)
	WW	0.8943(1.5e-4)	0.9644(1.8e-4)	0.9834(0)	0.9909(0)	0.9963(4.0e-5)
	SVM-k	0.8944(1.4e-4)	0.9641(3.0e-4)	0.9841(2.2e-4)	0.9913(4.3e-4)	0.9964(1.6e-4)
	CE-k	0.8950(1.2e-4)	0.9635(3.6e-4)	0.9842(1.1e-4)	0.9915(8.0e-5)	0.9962(4.9e-5)
	LDR-KL	**0.8961(0)**	0.9648(0)	0.9843(1.3e-4)	0.9918(5.5e-5)	0.9964(8.9e-5)
	LDR-k-KL	**0.8961(0)**	**0.9651(0)**	**0.9851(1.9e-4)**	**0.9922(2.5e-4)**	**0.9967(1.3e-4)**
CIFAR-100	CE	0.6664(7.5e-5)	0.7855(4.9e-5)	0.8401(1e-4)	0.8721(1.1e-4)	0.8917(0)
	CS	0.6652(2.7e-4)	0.7833(2.2e-4)	0.8381(4.8e-4)	0.8695(4.8e-4)	0.8895(1.5e-4)
	WW	0.6671(3.6e-4)	0.7868(4.7e-4)	0.8393(4.8e-4)	0.8725(4e-4)	0.8916(5.8e-4)
	SVM-k	0.6652(2.7e-4)	0.7849(6.0e-4)	0.8361(1.2e-3)	0.8710(1.1e-3)	0.8950(7.0e-4)
	CE-k	0.6664(7.5e-5)	0.7861(8e-4)	0.8409(4.3e-4)	0.8736(2.3e-4)	0.8957(3e-4)
	LDR-KL	**0.6683(7.5e-5)**	0.7883(4.9e-5)	**0.8430(7.5e-5)**	0.8730(8e-5)	0.8953(4.9e-5)
	LDR-k-KL	**0.6683(7.5e-5)**	**0.7885(2.7e-4)**	0.8424(3.2e-4)	**0.8783(2e-4)**	**0.8999(3e-4)**
ImageNet	CE	0.7471(3.5e-4)	0.8489(2.1e-4)	0.8872(2.6e-4)	0.9075(3.9e-4)	0.9205(3.3e-4)
	CS	0.7534(4.1e-4)	0.8506(5.6e-4)	0.8886(4.3e-4)	0.9077(2.7e-4)	0.9207(3.1e-4)
	WW	0.7170(8.9e-5)	0.8326(4.5e-5)	0.8771(4.6e-5)	0.9016(3.5e-5)	0.9179(5.7e-5)
	SVM-k	0.7534(4.1e-4)	0.8527(5.7e-4)	0.8915(4.7e-4)	0.9111(5.3e-4)	0.9248(3.9e-4)
	CE-k	0.7471(3.5e-4)	0.8499(3.6e-4)	0.8896(3.9e-4)	0.9111(6.8e-4)	0.9248(3.2e-4)
	LDR-KL	**0.7575(1.8e-4)**	**0.8568(1.3e-4)**	**0.8939(1.5e-4)**	0.9138(9.8e-5)	0.9270(6.2e-5)
	LDR-k-KL	**0.7575(1.8e-4)**	0.8551(2.2e-4)	0.8932(1.5e-4)	**0.9145(7.4e-05)**	**0.9273(6.2e-5)**

We grid-tune the hyper-parameters for all the baseline models by 5-fold cross validation. The ℓ_2-regularization hyper-parameters are tuned in the range {1e-4,1e-3,1e-2,0} for all the loss functions. The margin hyper-parameters $c(y, k)$ are tuned in the range {1e-1,1,1e1} for all the margin-based loss functions. To avoid heavily tuning cost, we first grid-tune LDR-KL and LDR-k-KL with $c(y, k)$ and distributional weight (DW) in the range {1e-1,1,1e1} with fixed ℓ_2-regularization hyper-parameter as 1e-4. After that, tune them with ℓ_2-regularization hyper-parameters in the same range as other baselines.

It is worth noting that the top-k losses for $k = 1$ (SVM-k, CE-k and LDR-k-KL) would be reduced to their classical top-1 version losses (CS, CE and LDR-KL). Hence, they have the same top-1 precision values as their top-1 version losses. We run each method on each dataset independently 5 times and report the averaged top-k (k=1,2,3,4,5) accuracy together with their standard deviations in the Table 2.

As we can see from the results, LDR-KL and LDR-k-KL performs better comparing to other existing loss functions. For aloi dataset with a large number of classes, LDR-KL and LDR-k-KL are consistently better than other methods, which means the proposed LDR loss could benefit from the data with a large class number. It is consistent with our assumption because it is more possible that such data could have more potential unlabeled true class. For news20 and letter datasets, the proposed

LDR-KL loss is generally better than CE loss for top-1 precision; and the proposed LDR-k-KL loss is generally better than CE-k loss. The non-smooth SVM losses are better for some cases thanks to the large margin classifier effect.

Image Daatasets: To eliminate the effect of different losses on feature learning (wich is not the focus of this paper), we conduct experiments by using a pre-trained network for extracting the feature vector of an image. We first pre-train deep neural networks on the benchmark datasets (Resnet-32 on cifar and Resnet-50 on imagenet by optimizing the CE loss) for extracting a feature representation [26]; then based on the pre-trained hidden features we learn linear models by optimizing different loss functions. For the pre-training process, we use momentum SGD with stage-wised decreasing step size to optimize 50 epochs for cifar-10, 150 epochs for cifar-100 and 90 epochs for Imagenet. The hidden feature dimension for CIFAR-(10 and 100) is 64; and the hidden feature dimension for Imagenet is 2048. For learning the linear model based on the extract feature representations, we also utilize momentum SGD with stage-wised decreasing step size (10000 iterations as one stage; initial learning rate is 0.1 and decreasing factor is 0.1) for optimizing different loss functions, with a total of 20000 and 50000 iterations for CIFAR-(10 and 100) and Imagenet, respectively. The batch size is 128 for all the experiments in this work. We keep the same protocol for model validation and evaluation as for the structured data, except that we keep the default testing splitting. The results are summarized in the Tables 2. From the results, LDR-KL and LDR-k-KL losses generally outperform the other loss functions. The advantages are more obvious for cifar100 and Imagenet, who have a large number of classes. For cifar10, the classes should be more clearly separated and easier for both human and machine learning model to recognize.