Optimizing Partial Area Under the Top-k Curve: Theory and Practice

Zitai Wang, Qianqian Xu*, Senior Member, IEEE, Zhiyong Yang, Yuan He, Xiaochun Cao, Senior Member, IEEE, and Qingming Huang*, Fellow, IEEE

Abstract—Top-k error has become a popular metric for large-scale classification benchmarks due to the inevitable semantic ambiguity among classes. Existing literature on top-k optimization generally focuses on the optimization method of the top-k objective, while ignoring the limitations of the metric itself. In this paper, we point out that the top-k objective lacks enough discrimination such that the induced predictions may give a totally irrelevant label a top rank. To fix this issue, we develop a novel metric named partial Area Under the top-k Curve (AUTKC). Theoretical analysis shows that AUTKC has a better discrimination ability, and its Bayes optimal score function could give a correct top-K ranking with respect to the conditional probability. This shows that AUTKC does not allow irrelevant labels to appear in the top list. Furthermore, we present an empirical surrogate risk minimization framework to optimize the proposed metric. Theoretically, we present (1) a sufficient condition for Fisher consistency of the Bayes optimal score function; (2) a generalization upper bound which is insensitive to the number of classes under a simple hyperparameter setting. Finally, the experimental results on four benchmark datasets validate the effectiveness of our proposed framework.

Index Terms—Machine Learning, Label Ambiguity, Top-k Error, AUTKC Optimization.

1 INTRODUCTION

Top-1 error, which checks whether the annotated label has the highest score, is a ubiquitous measure in classification tasks. Implicitly, this metric assumes that (a) the annotated label is exactly the ground-truth, i.e., (no label noise [2]), and (b) the ground-truth label is easy to discriminate from the others, i.e., (no label ambiguity [3], [4]). To ensure (a), we can either improve the data quality to clean the possible noises, or improve the quality of the training model to resist the adverse effect of noises. Different from the noises, label ambiguity is an intrinsic property of the label semantic. Consequently, it cannot be eliminated by simply improving the quality of data once the label space is fixed. In fact, (b) does not necessarily hold as the number of classes in benchmark datasets grows rapidly [1], [5]. As shown in Fig. 1, when facing numerous classes, there are two major issues that make the ground-truth label ambiguous. One is that the semantic similarity among classes inevitably emerges as the size of label space grows larger, leading to high error rates of the top-1 predicted labels even for humans [4]. The other is that, when the number of classes is larger, each instance becomes more likely to be relevant with multiple classes, making each relevant class a reasonable prediction. Dealing with label ambiguity is thus a harder problem than label noises. In this sense, our goal is to seek a better solution against label ambiguity in this paper.

Historically, there exist two solutions against label ambiguity. One solution is to perform multi-label learning [3], [6], where each instance is annotated with multiple labels to cover ambiguous labels as much as possible. The
other one is to relax the top-1 error and perform top-k optimization [4], [7], [8], [9], where ambiguous labels are allowed to be ranked higher than the ground-truth. In this way, one can approximate the multi-label setting with a single-label dataset. Although the first solution is natural and straightforward, multi-label annotation is generally intractable for large-scale datasets. In practice, single-label datasets [1], [3] enjoy much higher scalability than multi-label datasets [10], [11] in terms of both the number of classes (hundreds v.s. dozens) and the number of instances (million v.s. thousands). Considering this fact, this paper focuses on the second solution, with a special interest in directly optimizing the top-k measure.

In this direction, existing literature has proposed a series of methods to optimize the top-k objective [4], [7], [8], [9]. However, the limitations of the metric itself have not been well understood. In this paper, we argue that the top-k error lacks enough discrimination, leaving some fatal errors free from penalty. Specifically, when the top-k error allows ambiguous labels to be ranked higher, it also relaxes the punishment on irrelevant labels. As shown in Fig.1, even if an irrelevant label Dog is ranked higher than the ground-truth label Mountain and the ambiguous labels Valley and Sky, the top-k error can still be zero, as long as Mountain appears in the predicted top-k list. To fix this issue, this paper takes a further step toward a more discriminating metric.

First of all, in Sec.3.2, the analysis suggests that looking at the performance at a single fixed k is insufficient to ensure the true prediction quality. In view of this, a natural solution is to evaluate the performance from a comprehensive perspective. To this end, we design a new measure named partial Area Under the Top-k Curve (AUTKC) in Sec.3.3, as an analogy with AUC [12]. The comparison in Sec.3.4 reveals that AUTKC is strictly consistent and more discriminating than the top-k error, which helps it overcome the limitations of the top-k measure.

Considering its advantages, it is appealing to design learning algorithms that optimize AUTKC efficiently. To this end, we have the following abstract formulation of the optimization problem, whose details are shown in Eq.(9).

\[
\min_{f} \mathbb{E}_{x} \left[ \sum_{k} \ell_{k}(f; x) \right],
\]

where the loss \( \ell_{k} \) takes the model prediction \( f(x) \) as the input. As presented in Sec.4, the main challenges are two-fold: (C1) \( \ell_{k} \) is discrete and non-differentiable, making the objective difficult to optimize; (C2) the data distribution is generally unavailable, making it impossible to calculate the expectation.

Targeting (C1), one common strategy is to replace \( \ell_{k} \) with a differentiable surrogate loss. But, what kind of surrogate loss should we select? The basic requirement is Fisher consistency, that is, we should recover the Bayes optimal solution of Eq.(1). In view of this, in Sec.5.1, we present a necessary and sufficient property for the Bayes-optimal function induced by Eq.(1). It suggests that the Bayes optimal score function must preserve the ranking of the underlying top-K labels w.r.t. the conditional probability \( P(y = i | x) \). Since the conditional probability for the irrelevant labels cannot surpass the relevant ones, the optimal solution naturally removes the irrelevant classes from the top list. On top of this, we develop a sufficient condition for the Fisher consistency for AUTKC in Sec.5.2. Based on this condition, we find that common surrogate losses are consistent with AUTKC as long as the model outputs are bounded, including the square loss, the exponential loss and the logit loss. However, the hinge loss, which is the standard surrogate loss of the top-k optimization methods, is inconsistent.

In Sec.6, we further construct an empirical optimization framework for AUTKC to relieve (C2). Specifically, we turn to minimize the unbiased empirical estimation of the surrogate loss over the training set, thus avoiding the direct calculation of the expectation. A basic question is whether the performance on the training set can generalize well to the unseen data. To answer this question in the context of label ambiguity, we construct the upper bound of the generalization error in terms of the number of classes \( C \) and deep neural networks. The result shows that we can obtain a generalization bound that is insensitive to \( C \).

Finally, we conduct the experiments on four benchmark datasets in Sec.7, where the number of classes ranges from 10 to 365.

In summary, the contribution of this paper is three-fold:

- **New measure**: We provide a novel performance measure named AUTKC to handle the label ambiguity in large-scale classification.
- **Theoretical guarantee**: The proposed framework for AUTKC optimization is supported by consistent surrogate losses and a generalization bound insensitive to the number of classes.
- **Empirical validation**: The empirical results not only show the superiority of the proposed framework, but also validate the theoretical results.

## 2 RELATED WORK

### 2.1 TOP-k Optimization

The early literature on top-k optimization focuses on the binary scenarios where only the top-ranked instances are of interest, such as retrieval (relevant/irrelevant) [13] and recommend (like/dislike) [14]. These methods minimize a pairwise ranking loss that punishes more on the errors related to the top-ranked instances [15], [16], [17].

As the scale of classes grows rapidly in modern benchmark datasets [1], [5], the semantic overlap between classes inevitably emerges. In this light, the top-k error has been extended to multiclass classification [1], [18], [19], [20]. Similar to the binary case, this metric evaluates the model performance on the top-ranked items. The difference lies in it measuring the ranking of the ground-truth label, rather than that of instance. The early work in this direction adopts an alternative approach to minimize the top-k error. For example, [21] models the classes missing from the annotation by the structured learning technique with latent variables. [22] analyzes the problem from the perspective of submodular reward functions. [23] proposes a probabilistic method explicitly modeling the total number of objects in an image. Besides, [24], [25] use metric learning to find potential labels, especially when the number of classes is dynamic.
Recently, more efforts focus on directly minimizing the top-$k$ error. To this end, [7] relaxes the multiclass hinge loss to penalize the top-$k$ highest prediction. Although this approach is intuitive, the induced loss functions are not top-$k$ consistent [4]. In other words, minimizing these losses does not guarantee an ideal top-$k$ performance. To this end, [4] analyzes the calibration property of some top-$k$ loss functions truncated from the multiclass hinge loss and the cross-entropy loss. The theoretical results are further extended to multi-label classification [8]. Most recently, [9] points out that the condition for top-$k$ calibration proposed by [4] is invalid when ties exist in the prediction. To fix this issue, [9] develops a necessary and sufficient property for top-$k$ optimality named top-$k$ preserving. Based on this property, a top-$k$ consistent hinge-like loss is finally established.

Compared with existing efforts on the optimization method of the top-$k$ objective, this paper focuses on the limitations of the metric itself, whose details are shown in Sec.3.2. To solve this problem, we propose a novel metric named AUTKC. Moreover, an efficient framework for AUTKC optimization is also established.

### 2.2 Label Ranking

The setting we discuss is naturally associated with the label ranking problem, where the predictions are also evaluated by the ranking of labels. Specifically, given the label space $Y$, label ranking aims to assign each instance $x$ with the correct ranking of all the labels, that is, a complete/partial, transitive, and asymmetric relation $\succ$ defined on $Y$, where $i \succ j$ means that the label $i$ precedes the label $j$ in the ranking associated with $x$. According to the taxonomy established by [26], label ranking methods can be divided into four categories: the ones decomposing the original problem to multiple simple objectives such as pointwise function [27], [28] and pairwise ranking loss [29], [30]; probabilistic methods including tree-based model [31], [32], [33], [34], [35], Gaussian mixture model [36] and structured learning [37]; the ones based on similarity [38], [39], [40]; and the rule-based ones [41], [42]. Please refer to the surveys for more details [26],[43].

Compared with label ranking, top-$k$ optimization follows the setting of classification, where each instance $x$ is assigned with a single label $y \in Y$, i.e., the ground-truth label. As a result, the common measures used in label ranking become invalid, such as Spearman’s rank [44] and Kendall’s tau [45]. And top-$k$ optimization only considers the ranking associated with the ground-truth label, i.e., ensuring $y \succ x \iff j \in Y$. Concretely, top-$k$ optimization requires that $|\{j \in Y : j \succ x \}| \leq k$, where $|\cdot|$ represents the cardinality of a set. Besides, our analysis in Sec.5.1 shows that optimizing AUTKC encourages the prediction to preserve the ranking of the underlying top-$K$ labels. In this sense, AUTKC optimization essentially aims to recover the ranking of partial labels under a restricted condition where only the ground-truth label is available.

### 2.3 AUC Optimization

Our proposed framework is also related to AUC optimization, where the optimization objective is also a pairwise ranking loss. Specifically, AUC, the Area Under the Receiver operating characteristic (ROC) Curve, measures the probability that the positive instances are ranked higher than the negative ones, with the assumption that the possibility to observe ties in the comparisons equals zero [12]. Being insensitive to the label distribution, AUC has become a popular metric in the class-imbalanced applications as disease prediction [46] and rare event detection [47], [48], [49].

Traditional research on AUC optimization targets the whole area under the ROC curve. At the early stage, the major studies in this direction focus on the off-line setting [50], [51], [52], [53], [54]. As the scale of datasets increases rapidly, it becomes infeasible to optimize the pairwise loss in a full-batch manner. For this reason, more studies explore the extension of AUC optimization methods in the online setting [55], [56], [57], [58], [59], [60], [61], [62]. Besides, the theoretical guarantee of the learning framework is another appealing topic, including generalization bound [63], [64], [65], [66], [67] and consistency property [68], [69]. Recently, more literature focuses on the partial area under the ROC curve since only the order of the top-ranked instances are of interest in many practical applications [70], [71], [72], [73]. In this sense, partial AUC optimization is related to top-$k$ optimization in binary scenarios.

Compared with partial AUC optimization, our proposed framework focuses on the ranking of multiple labels. Moreover, we assume that ties exist in the scores. These two differences will lead to totally different theoretical results, whose details are shown in Sec. 5 and Sec. 6.

---

**TABLE 1**

Some important Notations used in this paper.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>the number of classes</td>
</tr>
<tr>
<td>$K$</td>
<td>the maximum number of relevant labels</td>
</tr>
<tr>
<td>$X, Y$</td>
<td>the input space and the label space</td>
</tr>
<tr>
<td>$D$</td>
<td>the joint distribution defined on $Z = X \times Y$</td>
</tr>
<tr>
<td>$\eta(x) y$</td>
<td>the conditional distribution $P [y</td>
</tr>
<tr>
<td>$f$</td>
<td>the score function mapping $X$ to $R^C$</td>
</tr>
<tr>
<td>$F$</td>
<td>the set of score functions</td>
</tr>
<tr>
<td>$L$</td>
<td>the loss function mapping $Y \times R^C$ to $R_+$</td>
</tr>
<tr>
<td>$\pi_s(y)$</td>
<td>the index of $s_y$ when sorting $s \in R^C$ descendingly</td>
</tr>
<tr>
<td>$s_{[k]}$</td>
<td>the $k$-th greatest entry of $s \in R^C$</td>
</tr>
<tr>
<td>$(s_{[y]} y)</td>
<td>k$</td>
</tr>
<tr>
<td>$\ell_{0-1}, \ell$</td>
<td>the 0-1 loss and its surrogate loss</td>
</tr>
<tr>
<td>$\mathcal{R}_K(f)$</td>
<td>the original AUTKC risk</td>
</tr>
<tr>
<td>$\mathcal{R}_K(f^*)$</td>
<td>the optimal AUTKC risk</td>
</tr>
<tr>
<td>$\mathcal{R}_K(f^*)$</td>
<td>the surrogate AUTKC risk</td>
</tr>
<tr>
<td>$S$</td>
<td>the dataset sampled from $D$</td>
</tr>
<tr>
<td>$\mathcal{R}_K(f, S)$</td>
<td>the empirical AUTKC risk</td>
</tr>
<tr>
<td>$\Theta(S, \delta)$</td>
<td>the empirical Gaussian Complexity</td>
</tr>
<tr>
<td>$\Phi(S, \delta)$</td>
<td>the generalization error of AUTKC</td>
</tr>
<tr>
<td>$\mathcal{C}(\Theta, \epsilon, d)$</td>
<td>covering number defined on the metric space $(\Theta, d)$</td>
</tr>
<tr>
<td>$\mathcal{F}(\beta, \nu)$</td>
<td>the set of neural networks we discuss in Sec.6.2.2</td>
</tr>
<tr>
<td>$\text{vec}(-)$</td>
<td>the vectorization operation</td>
</tr>
<tr>
<td>$mt(-)$</td>
<td>the operator matrix</td>
</tr>
</tbody>
</table>
measure for classification is top-of-the-score function with the only one guess, i.e.,

\[
\text{TOP-k}(f) = \mathbb{E}_{x \sim \mathcal{D}} \left[ \text{err}_k(y, f(x)) \right],
\]

where

\[
\text{err}_k(y, f(x)) = \mathbf{1} \left[ \pi_f(x)(y) > k \right],
\]

1 \[\cdot\] is the indicator function, and \(\pi_f(x)(y) \in \{1, \cdots, C\}\) denotes the index of \(f(x)_y\) when we sort \(f(x)\) in a descending order. Meanwhile, top-k accuracy, denoted as \(\text{TOP-k}^\top\), is defined as \(1 - \text{TOP-k}^\top\); thus the two metrics are equivalent.

In the following discussion, we use TOP-k to refer to top-k error and accuracy when there exists no ambiguity.

Note that different ways of breaking ties will lead to different \(\pi_f(x)(y)\). Traditional methods assume that no ties exist since the number of classes is relatively small. However, the problem becomes non-negligible when calculating TOP-k since label ambiguity generally involves a large number of classes. Following the prior arts on TOP-k optimization [9], we adopt the worst-case assumption:

**Assumption 1.** There exist ties in \(f(x)\), and all the ties will be wrongly broken. In other words, we have the following formulations:

- Given \(y_1, y_2\) such that \(\eta(x)_{y_1} > \eta(x)_{y_2}\), if \(f(x)_{y_1} = f(x)_{y_2}\), then we have \(\pi_f(x)(y_1) > \pi_f(x)(y_2)\).
- Given the ground-truth label \(y\), for any \(y' \neq y\) such that \(f(x)_{y'} = f(x)_y\), we have \(\pi_f(x)(y) > \pi_f(x)(y')\).

**Remark 1.** The two formulations apply to different situations:

- The first formulation is a population-level assumption where the conditional probability \(\eta(x)\) is available. It states that, when ties exist, the ranking results of \(f(x)\) are opposite to those of \(\eta(x)\). We apply this formulation in the proof of Thm.3, Thm.4, and Thm.5.
- The second formulation is an empirical-level assumption where only the ground-truth label is available. It states that, when ties exist, the ground-truth label will be ranked lower. We apply this formulation in the proof of Thm.2.
- In summary, these two conditions ensure that we will always choose the worst result when there is a tie in a conservative sense.

### 3.2 Motivation

As mentioned in the introduction, TOP-k has its limitations. In this subsection, we will make a further analysis to reveal its reasons.

On one hand, TOP-k is not discriminating enough, leaving some fatal errors free from penalty (L1). As shown in Fig.3, we assume that Lake is the ground-truth label; Sky and Cloud are the labels with ambiguity; Beach and...
Valley are the irrelevant labels. Consider the following two score function such that:

\[ f_1: \text{Lake} \succ \text{Sky} \succ \text{Cloud} \succ \text{Beach} \succ \text{Valley}, \]
\[ f_2: \text{Beach} \succ \text{Lake} \succ \text{Sky} \succ \text{Cloud} \succ \text{Valley}. \]

Obviously, \( f_1 \) is an ideal score function since it ranks Lake first, while \( f_2 \) is a bad score function since it ranks the irrelevant label Beach higher than the ground-truth label.

Based on this observation, we expect

\[ \text{err}(y, f_2(x)) > \text{err}(y, f_1(x)). \]

However, when \( k \geq 2 \), we have

\[ \text{err}_k(y, f_1(x)) = \text{err}_k(y, f_2(x)) = 0. \]

In other words, although using the top-\( k \) error can relax the punishment on label ambiguity, it also relaxes the punishment on the fatal errors. In this sense, we need to seek out a more discriminating metric to selectively relax the punishment.

One may wonder whether a clever selection of \( k \) could eliminate this limitation. Unfortunately, the number of ambiguous labels generally differs among classes, while the hyperparameter \( k \) is universal for the classes. This contradiction leads to a dilemma during the training period: if the value of \( k \) is too small for some classes, label ambiguity might make these classes hard to optimize, which is the motivation of TOP-k optimization. But, if the value of \( k \) is too large, irrelevant labels will be more likely to be ranked higher than the ground-truth label, which is also unexpected. This suggests that merely looking at a single \( k \) is not enough to address (L1).

On the other hand, TOP-k is a static measure, ignoring the changes in operator's decision condition (L2). To be specific, empirical results suggest that models trained with different hyperparameters \( k \) generally perform inconsistently on the top-\( k \) curve \([4, 8]\). As shown in Fig. 4, model \( f \) outperforms model \( g \) when \( k < 4 \), while the results reverse when \( k > 4 \). In view of this, the operator might ask: which model should we select to deploy? In practice, the answer depends on which \( k \) is of interest during the deployment phase. However, TOP-k optimization methods fix the value of \( k \) during the training phase. As a result, the model performance cannot be guaranteed if the \( k \) of interest changes after the model is deployed.

Being aware of the limitations of TOP-k, we have the following question:

(Q1) Whether there exists a "better" metric for label ambiguity?

3.3 AUTKC and its Partial Variant

Our answer to (Q1) is to evaluate the score function from a more comprehensive perspective, rather than a single point of the top-\( k \)-curve. Taking the inspiration from Area Under the ROC Curve (AUC) \([12]\), we propose to adopt the area under the top-\( k \)-curve (AUTKC) to evaluate the quality of the given score function:

\[
\text{AUTKC-W}(f) := \mathbb{E}_{z \sim D} \left[ \frac{1}{C} \int_1^{C} \text{err}_k(y, f(x)) dk \right],
\]

where \( K_1, K_2 \) are the minimum and the maximum number of ambiguous labels, respectively. Since TOP-1 is the most common metric of interest, and there generally exist some instances involving no ambiguous labels, we set \( K_1 = 1 \) and denote \( K_2 \) as \( K \) for the sake of conciseness in the following discussion.

Furthermore, since \( k \) is discrete, Eq. (5) enjoys a much simpler formulation:

\[
\text{AUTKC-P}(f) := \mathbb{E}_{z \sim D} \left[ aerr_K(y, f(x)) \right],
\]

where

\[
aerr_K(y, f(x)) = \frac{1}{K} \sum_{k=1}^{K} \text{err}_k(y, f(x)).
\]

Note that we denote AUTKC-P as AUTKC in the following discussion for the sake of convenience. Similar to TOP-k, the partial area under top-\( k \)-accuracy curve, denoted as AUTKC, is equivalent to AUTKC, and we use AUTKC to refer to the two metrics when there exists no ambiguity.

3.4 AUTKC vs TOP-k

Intuitively, AUTKC is a better measure than TOP-k. On one hand, AUTKC is more discriminating than TOP-k, which helps overcome the limitation (L1). For example, when we set \( K = 3 \), AUTKC tells us \( f_1 \) is the ideal score function and indeed better than \( f_2 \):

\[ aerr_K(y, f_1(x)) = \frac{1}{3} (0 + 0 + 0) = 0, \]

and

\[ aerr_K(y, f_2(x)) = \frac{1}{3} (1 + 0 + 0) = \frac{1}{3}. \]
On the other hand, AUTKC only requires a universal hyperparameter $K$ for all the instances, that is, the maximum number of ambiguous labels. Such a simplification is beneficial to overcoming the limitation \((L2)\).

Although the aforementioned example indicates the advantages of AUTKC, more theoretical clues are necessary to validate our argument. To this end, we first present the definition of consistent and discriminating, which is a classic technique for comparing measures [12].

Definition 1 (Degree of consistency [12]). Given two measures $f$ and $g$ on domain $\Psi$ and two predictions $a, b \in \Psi$, let $R = \{(a, b) | f(a) > f(b), g(a) > g(b)\}, S = \{(a, b) | f(a) > f(b), g(a) < g(b)\}$. The degree of consistency between $f$ and $g$ is defined as $C = \frac{|R|}{|R| + |S|}$, where $|\cdot|$ is the cardinality of a set.

Remark 2. The degree of consistency estimates the probability that the two measures make the same judgment on the given domain. A relatively high degree of consistency is a necessary condition for measure comparison.

Definition 2 (Degree of discrimination [12]). Given two measures $f$ and $g$ on domain $\Psi$ and two predictions $a, b \in \Psi$, let $P = \{(a, b) | f(a) > f(b), g(a) = g(b)\}, S = \{(a, b) | g(a) > g(b), f(a) = f(b)\}$. The degree of discrimination between $f$ and $g$ is $D = \frac{|P|}{|S|}$.

Remark 3. Degree of discrimination describes whether a measure could find the discrepancy that the other measure fails to discover. For example, the top-1 error is intuitively more discriminating than the top-$k$ error when $k > 1$.

Definition 3 (Consistent and discriminating [12]). The measure $f$ is consistent and more discriminating than $g$ if and only if $C > 0.5$ and $D > 1$. In this case, we conclude, intuitively, that $f$ is a “better” measure than $g$. Especially, if $C = 1.0$ and $D = \infty$, we say the measure $f$ is strictly consistent and more discriminating than $g$.

On top of consistency and discrimination, the following theorem suggests that AUTKC is a better measure than TOP-k. Please see Appendix A for the proof.

Theorem 1. For any $k < K$, AUTKC is strictly consistent and more discriminating than TOP-k.

Remark 4. Being strictly consistent with TOP-k for any $k < K$ means that AUTKC is insensitive to the $k$ of interest, which overcomes the limitation \((L2)\). Meanwhile, being more discriminating implies that AUTKC might be a better optimization objective than TOP-k, thus providing a clue to overcoming the limitation \((L1)\). Further analysis in Sec. 5.1 will support these conjectures.

4 Roadmap for AUTKC Optimization

So far, we have known the advantage of AUTKC. Then, a natural question is

(Q2) How to design a learning algorithm that optimizes AUTKC effectively?

Following the standard machine learning paradigm [74], we first reformulate the metric optimization problem to a risk minimization problem:

\[
\begin{align*}
(\text{OP}_0) \quad & \min_{f} \mathbb{E}_{z \sim D} \left[ \frac{1}{K} \sum_{k \leq K} \ell_{0-1} (s_y - (s_{\setminus y})_k) \right] \\
& = \mathbb{E}_{z \sim D} \left[ \frac{1}{K} \sum_{k \leq K} \ell_{0-1} (s_y - (s_{\setminus y})_k) \right], \\
(\text{OP}_1) \quad & \min_{f} \mathbb{E}_{z \sim D} \left[ \frac{1}{K} \sum_{k=1}^{K+1} \ell_{0-1} (s_y - s_k) \right], \\
& \text{where } s_y := f(x)_y \text{ denotes the score of the class } y, (s_{\setminus y})_k \text{ is the } k\text{-th greatest entry of } f(x) \text{ except } s_y, \text{ and } \ell_{0-1}(l) := 1 \text{ if } l \leq 0 \text{ is the 0-1 loss. Note that we set } \ell_{0-1}(1) = 1 \text{ since according to Asm.1, any ties will lead to an error.}
\end{align*}
\]

For the sake of conciseness, we provide an equivalent reformulation of \((\text{OP}_0)\), where the annoying operation \(\setminus y\) is removed. See Appendix B for the proof.

Theorem 2. When Asm.1 holds, the original optimization problem \((\text{OP}_0)\) is equivalent to the following problem:

\[
(\text{OP}_1) \quad \min_{f} \mathbb{E}_{z \sim D} \left[ \frac{1}{K} \sum_{k=1}^{K+1} \ell_{0-1} (s_y - s_k) \right].
\]

Remark 5. Thm. 2 bridges AUTKC optimization with two classic problems: AUC optimization and label ranking. Specifically, AUC optimization is a binary classification problem with a pairwise ranking objective [12]:

\[
(\text{OP}_\text{AUC}) \quad \min_{f, D^+, D^-} \mathbb{E}_{x^+, x^-, y, D^+, D^-} \left[ \ell_{0-1} (f(x^+) - f(x^-)) \right],
\]

where $D^+, D^-$ are the distribution of positive and negative samples, respectively. If we view the ground-truth label as the positive instance and the other labels as the negative ones, then AUTKC optimization is similar to a partial variant of AUC optimization in the context of label ranking. Although partial AUC optimization has been well studied [70], [71], [72], [73], there still exist two main differences:

- The objective of AUC optimization assumes that there exist no ties in $f(x)$, which contradicts Asm.1.
- The expectation operation of AUC involves two samples, while in AUTKC it only involves one label.

These differences will lead to different theoretical results and practical implementation. Please refer to Sec. 5 and Sec. 6 for the details.

According to Eq.(9), the main challenges for directly minimizing $\mathcal{R}_K(f)$ are two-fold:

\begin{itemize}
  \item [(C1)] The loss function $\ell_{0-1}$ is not differentiable;
  \item [(C2)] The data distribution $D$ is unavailable.
\end{itemize}

Next, we will present the solutions to (C1) and (C2) in Sec. 5 and Sec. 6, respectively. Finally, we will obtain an end-to-end framework that optimizes the empirical risk of AUTKC efficiently.

5 Surrogate Risk Minimization for AUTKC

In this section, our task is to handle (C1). Specifically, the discrete $\ell_{0-1}$ in \((\text{OP}_1)\) will be replaced by a differentiable
surrogate loss \( \ell \), leading to the following surrogate risk minimization problem:

\[
(\text{OP}_2) \quad \min_f R_{K}^\ell(f) := \mathbb{E}_{x \sim D} \left[ \frac{1}{K} \sum_{k=1}^{K+1} \ell(s_{y} - s_{[k]}) \right]. \tag{10}
\]

But,

\[(Q2-1) \quad \text{What kind of surrogate loss should we select?}\]

We next answer this question in three steps. First, we study the Bayes optimal function for AUTKC, which is the best score function we could approximate:

**Definition 4 (AUTKC Bayes optimal).** The score function \( f^* : X \to \mathbb{R}^C \) is AUTKC Bayes optimal if

\[
f^* \in \arg \inf_f R_{K}(f).
\]

Our goal is to find the solution to Eq.(11). To this end, we first present the top-\( K \) ranking-preserving property:

**Definition 5 (Top-\( K \) ranking-preserving property).** Given \( a, b \in \mathbb{R}^C \), we say that \( b \) is top-\( K \) ranking-preserving with respect to \( a \), denoted as \( \text{RP}_K(b, a) \), if for any \( k \leq K \),

\[
\pi_{b}(y) = k \implies \pi_{a}(y) = k.
\]

The negation of this statement is \( \neg\text{RP}_K(b, a) \). Note that \( \text{RP}_K \) is transitive. In other words, if \( \text{RP}_K(c, b) \) and \( \text{RP}_K(b, a) \), then we have \( \text{RP}_K(c, a) \).

Then, the following theorem illuminates the connection between \( \text{RP}_K \) and AUTKC optimality. Please see Appendix.C.1 for the proof.

**Theorem 3.** The score function \( f : X \to \mathbb{R}^C \) is AUTKC Bayes optimal if and only if the score is top-\( K \) ranking-preserving with respect to the conditional risk. More concretely, we have

\[
f(x)_{y_1} > f(x)_{y_2} \quad \text{if} \quad \eta(x)_{y_1} > \eta(x)_{y_2} \quad \text{for any} \quad k \leq K.
\]

It is worth noting that this theorem sheds more light on the advantages of optimizing AUTKC:

**Remark 6.** Top-\( K \) ranking-preserving is stricter than the optimality property for TOP-\( K \) optimization [9]. To be specific, given any \( x \), the top-\( K \) optimality property requires that

\[
\begin{align*}
&f(x)_{y} > f(x)_{y+1} \quad \text{if} \quad \eta(x)_{y} > \eta(x)_{y+1}, \\
&f(x)_{y} < f(x)_{y+1} \quad \text{if} \quad \eta(x)_{y} < \eta(x)_{y+1}.
\end{align*}
\]

In other words, this property only considers the ranking involving \( \eta(x)_{y} \) and \( \eta(x)_{y+1} \). As a comparison, Top-\( K \) ranking-preserving \( \text{RP}_K \) also considers the ranking between \( f(x)_{y_1} \) and \( f(x)_{y_2} \), where \( \eta(x)_{y_1} > \eta(x)_{y_2} > \eta(x)_{y} \). As shown in Fig.5, AUTKC optimality requires that

\[
\text{Lake} \succ \text{Sky} \succ \text{Cloud} \succ \text{Beach} \succ \text{others},
\]

while TOP-\( K \) optimality only requires that

\[
\text{Lake}, \text{Sky, Cloud, Beach} \succ \text{others}.
\]

The advantage of this strict property is three-fold:

- It reveals how AUTKC overcomes limitation \( (L1) \): optimizing \( \text{AUTKC} \) encourages the classes with a higher conditional probability being ranked higher, while the TOP-\( K \) objective does not.
- It reveals how AUTKC overcomes limitation \( (L2) \) and escapes from the dilemma that the value of hyperparameter \( K \) is hard to determine: as long as the hyperparameter \( K \) is not so small, the induced optimal prediction will rank the ground-truth label and the ambiguous ones higher than the irrelevant ones.
- It helps us obtain a simpler criterion to select surrogate loss, whose details are presented in Sec.5.2.

**Remark 7.** If a score function \( f \) is AUTKC Bayes optimal, no ties will exist in

\[
\{ f(x)_{[1]}, \ldots, f(x)_{[K]}, f(x)_{[K+1]} \}.
\]

In other words, optimizing \( \text{AUTKC} \) encourages breaking the ties, which is an appealing property. Meanwhile, ties might exist in

\[
\{ f(x)_{[K+1]}, \ldots, f(x)_{[C]} \}
\]

since it does not break the top-\( K \) ranking-preserving property.

### 5.2 Consistency analysis for surrogate risk minimization

On top of Thm.3, we could investigate the answer to the question \((Q2-1)\). A basic requirement for \( \ell \) is consistency. That is, the solution returned by minimizing \( R_{K}^\ell(f) \) should recover the Bayes optimal score function:

**Definition 6 (AUTKC consistency).** The surrogate loss \( \ell : \mathbb{R} \to \mathbb{R}_+ \) is consistent with AUTKC if for every function sequence \( \{f_t\}_{t=1,2,\ldots} \), we have:

\[
R_{K}^\ell(f_t) \to \inf_{f} R_{K}(f) \Rightarrow R_{K}(f_t) \to \inf_{f} R_{K}(f). \tag{12}
\]

Next, we provide a sufficient condition for AUTKC consistency, which enjoys a simpler formulation than checking whether the optimal solutions of \((\text{OP}_2)\) are Bayes optimal. Please see Appendix.C.2 for the proof.

**Theorem 4.** The surrogate loss \( \ell(t) \) is AUTKC consistent if it is bounded, differentiable and strictly decreasing.
So far, we have discussed the consistency w.r.t. all the measurable functions. However, common surrogate losses are not bounded. To this end, we next restrict the functions within a special function set $\mathcal{F}$, which induces the concept of $\mathcal{F}$-consistency:

**Definition 7 (AUTKC $\mathcal{F}$-consistency).** The surrogate loss $\ell : \mathbb{R} \rightarrow \mathbb{R}_+$ is consistent with AUTKC if for every function sequence $\{f_i\}_{i=1,2,\ldots} \in \mathcal{F}$, we have:

$$\mathcal{R}_K \ell (f_i) \rightarrow \inf_{f \in \mathcal{F}} \mathcal{R}_K \ell (f) \Rightarrow \mathcal{R}_K (f_i) \rightarrow \inf_{f \in \mathcal{F}} \mathcal{R}_K (f). \quad (13)$$

According to Thm.4, the following proposition and corollary clearly hold:

**Proposition 1.** Let $\mathcal{F}$ denote the set of functions whose outputs are bounded in $[0, 1]$. Then the surrogate loss $\ell(t)$ is $\mathcal{F}$-consistent with AUTKC if it is differentiable and strictly decreasing in $[0, 1]$.

**Corollary 1.** Let $\mathcal{F}$ denote the set of functions whose outputs are bounded in $[0, 1]$. Then the following statements hold:

- The square loss $\ell_{sq}(t) = (1 - t)^2$ is $\mathcal{F}$-consistent with AUTKC.
- The exponential loss $\ell_{exp}(t) = \exp(-t)$ is $\mathcal{F}$-consistent with AUTKC.
- The logit loss $\ell_{logit}(t) = \log(1 + \exp(-t))$ is $\mathcal{F}$-consistent with AUTKC.

Traditional TOP-k optimization methods adopt the hinge loss $\ell_{hinge}(t) = [1 - t]_+$ as the surrogate loss, where $[t]_+ = \max\{0, t\}$. However, we point out that its variant in AUTKC optimization is inconsistent, whose proof is shown in Appendix.C.3.

**Theorem 5.** The hinge loss is inconsistent with AUTKC when there exists $x$ such that $\sum_{k=K+2}^{C} \eta(x[i]) > \frac{K}{K+1}$.

**Proof Sketch.** Given any sample $x$ satisfying the above condition and its Bayes optimal score $s^*$, Thm.5 is induced by the fact that we could construct a score $s$ such that $-\mathcal{R}^\ell_P (s, \eta(x))$ but $s$ has a smaller risk than $s^*$ does. This construction holds by the property that $\ell_{hinge}(t) = 0$ for any $t \geq 1$. Please refer to Appendix.C.3 for the details.

### 6 Generalization Analysis for AUTKC Optimization

In this section, we aim to relieve (C2). Since the data distribution $\mathcal{D}$ is unknown, we turn to optimizing its empirical estimation based on the given dataset $\mathcal{S} = \{z^{(i)}\}_{i=1}^n$ sampled i.i.d. from $\mathcal{D}$. Let $s^{(i)}_y$ denote $f(x^{(i)})^{(y)}$ for convenience. Then, we have the following optimization problem:

$$(OP_3) \min_{f \in \mathcal{F}} \mathcal{R}_K \ell (f, \mathcal{S}) := \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{K+1} \ell \left( s^{(i)}_y - s^{(i)}_k \right), \quad (14)$$

The results in Sec. 5 guarantee a small AUTKC risk $\mathcal{R}_K$ when we minimize the surrogate risk $\mathcal{R}_K \ell$, as long as the surrogate loss $\ell$ is AUTKC consistent. But,

**(Q2-2)** Does minimizing the empirical risk $\mathcal{R}_K \ell (f, \mathcal{S})$ lead to a small generalization error $\mathcal{R}_K (f)$?

In other words, we should ensure $\mathcal{R}_K (f) - \mathcal{R}_K (f, \mathcal{S})$ is bounded with high probability, and the upper bound goes to zero as the number of samples increases, especially when the number of classes is large. To this end, we next provide a Gaussian-complexity-based analysis for this bound in Sec. 6.1. Since deep neural networks have been one of the most popular function sets for large-scale classification, we further present its result in Sec. 6.2.

#### 6.1 General Result for AUTKC Generalization Bound

**Gaussian Complexity** [74] is a classic measure of generalization error. Here we present its formal definition.

**Definition 8 (Gaussian Complexity).** Given the dataset $\mathcal{S} = \{z^{(i)}\}_{i=1}^n$, let $\mathcal{F}$ be a family of functions defined on the sample space $\mathcal{Z}$. Then, the empirical Gaussian complexity is defined as:

$$\phi_{\mathcal{S}} (\mathcal{F}) := \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n g^{(i)} f (z^{(i)}) \right], \quad (15)$$

where $g = (g^{(1)}, g^{(2)}, \ldots, g^{(n)})$ are the independent random variables sampled from the standard normal distribution $N(0,1)$.

Then, the following lemma provides the upper bound of $\mathcal{R}_K (f) - \mathcal{R}_K (f, \mathcal{S})$ via Gaussian complexity.

**Lemma 1 (Generalization Bound with Gaussian Complexity)** [74]. Suppose that the loss function $L$ is bounded by $B > 0$. Then, for any $\delta \in (0, 1)$, with a probability of at least $1 - \delta$, the following holds

$$\Phi(S, \delta) \leq \sqrt{2\pi} \phi_{\mathcal{S}} (L \circ \mathcal{F}), \quad (16)$$

where

$$\Phi(S, \delta) = \sup_{f \in \mathcal{F}} \left( \mathcal{R}_K \ell (f) - \mathcal{R}_K \ell (f, \mathcal{S}) \right) - 3B \sqrt{\frac{\log \frac{2}{\delta}}{2n}},$$

and $L \circ \mathcal{F} := \{L \circ f | f \in \mathcal{F}\}$ is the set of functions $L : \mathbb{R}_+ \rightarrow \mathbb{R}$ and the score function $f \in \mathcal{F}$.

In multi-class classification, this classic bound scales w.r.t. $O(\sqrt{C})$ [74], [75], [76]. This result is rather unfavorable when facing numerous classes. Thanks to the recent advance in learning theory, a Lipschitz assumption on the loss function $L$ can help eliminate this problem [77]:

**Assumption 2 (Lipschitz Continuity w.r.t. a Variant of the $\ell_2$-Norm)** [77]. The loss function $L : \mathbb{R}^C \rightarrow \mathbb{R}$ is Lipschitz continuous w.r.t. a variant of the $\ell_2$-norm involving index $y \in \mathcal{Y}$ if for all $s, s' \in \mathbb{R}^C$

$$|L(s) - L(s')| \leq L_1 \|s_1 - s'_1, \ldots, s_C - s'_C\|_2 + L_2 |s_y - s'_y|,$$

where $L_1, L_2$ are Lipschitz constants.

Then, according to Lemma 1 in [77], we obtain the following generalization bound:

**Proposition 2 (Generalization Bound with the Lipschitz continuity variant of the $\ell_2$-Norm).** If the loss function $L$ satisfies Assm.2, then for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$, we have

$$\Phi(S, \delta) \leq 2\sqrt{\pi} \left( L_1 C \phi_{\mathcal{S}} (\mathcal{F}) + L_2 \phi_{\mathcal{S}} (\mathcal{F}) \right). \quad (17)$$
where

\[
\tilde{S} = \left\{ (\mathbf{x}^{(1)}, 1), (\mathbf{x}^{(1)}, 2), \ldots, (\mathbf{x}^{(1)}, C) \right\},
\]

induced by \( \mathbf{w}^{(1)} \)

\[
(\mathbf{x}^{(2)}, 1), (\mathbf{x}^{(2)}, 2), \ldots, (\mathbf{x}^{(2)}, C) \right\},
\]

induced by \( \mathbf{w}^{(2)} \)

\[
(\mathbf{x}^{(n)}, 1), (\mathbf{x}^{(n)}, 2), \ldots, (\mathbf{x}^{(n)}, C) \right\},
\]

induced by \( \mathbf{w}^{(n)} \)

\[
\tilde{F} = \left\{ \tilde{f}(\mathbf{x}, y) := e^T_y f(\mathbf{x}) \mid f \in \mathcal{F}, \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y} \right\},
\]

and \( e_y = \left(0, 0, 1, 0, \cdots, 0 \right)^T \).

Note that this formulation generalizes the Thm.2 in [77], where \( f \) is restricted in linear models. Such generalization is necessary when we analyze the bound of neural networks.

According to Proposition 2, we need to present the Lipschitz property of the surrogate losses mentioned in Corollary 1. To this end, we first show the following theorem, whose proof is shown in Appendix.D.1.

**Theorem 6.** Assume that dom \( s_i \subset [0, B_i] \) for any \( i = 1, \cdots, C \), if \( \ell(t) \) is strictly decreasing and \( L_\ell \)-Lipschitz continuous in the range \([-B_i, B_i]\), then

\[
L_K(s, y) = \frac{1}{K} \sum_{k=1}^{K+1} \ell \left( s_y - s_{[k]} \right)
\]

is Lipschitz continuous w.r.t. \( f \) with Lipschitz constant pair \((\frac{L_{\ell} + \sqrt{2(K+1)}}{K}, \frac{L_{\ell} + \sqrt{2(K+1)}}{K})\).

On top of Thm.6, the following results are clear. We leave the rest of the proof in Appendix.D.1.

**Corollary 2.** Let \( \mathfrak{s} \) denote the normalized score \( \text{soft}(s) \), where soft is the softmax function. Then, the following statements hold:

- The square loss

\[
L_{K}^2(s, y) = \frac{1}{K} \sum_{k=1}^{K+1} \ell_{sq} \left( s_y - s_{[k]} \right)
\]

is Lipschitz continuous w.r.t. \( f \) with Lipschitz constant pair \((\frac{2\sqrt{2(K+1)}}{K}, \frac{2\sqrt{2(K+1)}}{K})\).

- The exponential loss

\[
L_{K}^{exp}(s, y) = \frac{1}{K} \sum_{k=1}^{K+1} \ell_{exp} \left( s_y - s_{[k]} \right)
\]

is Lipschitz continuous w.r.t. \( f \) with Lipschitz constant pair \((\frac{e\sqrt{2(K+1)}}{2K}, \frac{e\sqrt{2(K+1)}}{2K})\).

- The logit loss

\[
L_{K}^{logit}(s, y) = \frac{1}{K} \sum_{k=1}^{K+1} \ell_{logit} \left( s_y - s_{[k]} \right)
\]

is Lipschitz continuous w.r.t. a variant of the \( \ell_2 \)-norm with Lipschitz constant pair \((\frac{\sqrt{2(K+1)}}{2eK}, \frac{\sqrt{2(K+1)}}{2eK})\).

**6.2 Generalization Bound for Convolutional Neural Networks**

According to Proposition 2 and Thm.6, we could obtain the final generalization bound as soon as a proper upper bound of Gaussian complexity is available. To this end, we first present an important technique for the bound of Gaussian complexity named chaining in Sec.6.2.1. Then, the final result of convolutional neural networks is available under the given regularization condition in Sec.6.2.2.

**6.2.1 Chaining Bound for Gaussian Complexity**

Similar to Gaussian complexity, covering number [78] is another technique to measure the complexity of the given (pseudo)metric space:

**Definition 9** (\( \varepsilon \)-covering). Let \( \{ \mathcal{F}, d \} \) be a (pseudo)metric space and \( B(f, r) \) denote the ball of radius \( r \) centered at \( f \in \mathcal{F} \). Given \( \Theta \subset \mathcal{F} \), we say \( \mathcal{F}_m = \{ f_i \}_{i=1}^{m} \) is \( \varepsilon \)-covering of \( \Theta \) if \( \Theta \subset \bigcup_{i=1}^{m} B(f_i, \varepsilon) \). In other words, for any \( \theta \in \Theta \), there exists \( f_i \in \mathcal{F}_m \) such that \( d(\theta, f_i) < \varepsilon \).

**Definition 10** (Covering Number). On top of Def. 9, covering number of \( \Theta \) is defined as the minimum cardinality of any \( \varepsilon \)-covering of \( \Theta \):

\[
\mathcal{C}(\Theta, \varepsilon, d) := \min \{ m : \mathcal{F}_m \text{ is } \varepsilon \text{-covering of } \Theta \}.
\]

Our task is to bound the Gaussian Complexity with covering number. To this end, we first check the sub-Gaussian property of Gaussian Complexity. See Appendix.D.2 for the proof.

**Definition 11** (Sub-Gaussian Process w.r.t. the Metric \( d_X \)). A collection of zero-mean random variables \( \{ X_f, f \in \mathcal{F} \} \) is a sub-Gaussian process w.r.t. a metric \( d_X \) on \( \mathcal{F} \) if

\[
\mathbb{E}[e^{\lambda (X_f - X_{f'})}] \leq e^{\frac{\lambda^2 d_X^2(f, f')}{2}}, \forall f, f' \in \mathcal{F}, \lambda \in \mathbb{R}.
\]

**Proposition 3** (Sub-Gaussian Property of Gaussian Complexity). For any \( f \in \mathcal{F} \), let

\[
T_f(g) = \sum_{i=1}^{C} \sum_{j=1}^{C} g_{i,j} s_{j}^{(i)},
\]

Then, \( Z := \frac{1}{\sqrt{nC}} (T_f(g) - T_f(g')) \) is a sub-Gaussian process with metric

\[
d_{\infty,S}(s, s') = \max_{s \in S} |s_y - s'_y|,
\]

where \( s_{j}^{(i)} = f(\mathbf{x}_{j}^{(i)}) \).

Since \( d_{\infty,S}(s, s') \leq 1 \) and

\[
\mathcal{E}_S(\mathcal{F}) = \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{nC} T_f(g) \right],
\]

we have the chaining bound of Gaussian Complexity according to Dudley’s entropy integral [78]:
Proposition 4 (Chaining bound of Gaussian Complexity).
\[
\Theta_S(F) \leq \inf_{\alpha \geq 0} \left( C_1 \alpha + \frac{C_2}{nC} \int_0^1 \sqrt{\log \mathcal{G}(F, \epsilon, d_{\infty, S})} \, d\epsilon \right),
\]
(19)
where \( C_1, C_2 \) are two constants.

6.2.2 Practical Result for Convolutional Neural Networks

On top of Eq.(19), we are now ready to present the generalization bound of a family of neural networks, where both convolution layers and fully-connected layers exist. In fact, it is an application of the recent advance in [79].

The family of neural networks. The network consists of \( L_c \) convolutional layers and followed by \( L_f \) fully-connected layers. Each convolutional layer consists of a convolution followed by an activation function and an optional pooling operation. All the convolutions use zero-padding [80] with the kernel \( K N(l) \in \mathbb{R}^{p_l \times p_l \times c_{l-1} \times c_l} \) for layer \( l \in \{1, \ldots, L_c\} \). All the activation functions and pooling operations are 1-Lipschitz continuous. Meanwhile, we denote the parameters of the \( l \)-the fully-connected layer as \( V(l) \). Finally, let \( \Theta = \{K N(l), \ldots, K N(L_c), V(1), \ldots, V(L_f)\} \) denote the set of all the parameters in the given network.

Regularization of the parameters. The model input \( x \in \mathbb{R}^{d \times d \times c} \) satisfies \( \|\text{vec}(x)\| \leq c \), where \( c \) is the number of channels and \( \text{vec}(\cdot) \) is the vectorization operation. Note that convolution is essentially a linear operation, and we denote the operator matrix of the kernel \( K N(l) \) as \( m(K N(l)) \). We assume that the initial parameters, denoted as \( \Theta_0 \), satisfy

\[
\|m(K N_0(l))\|_2 \leq 1 + \nu, l = 1, \ldots, L_c,
\]
\[
\|m(V_0(l))\|_2 \leq 1 + \nu, l = 1, \ldots, L_f.
\]

And the distance from \( \Theta_0 \) to the current parameters \( \Theta \) is bounded:

\[
\|\Theta - \Theta_0\| := \sum_{l=1}^{L} \|m(K N_l) - m(K N_0(l))\|_2 + \sum_{l=1}^{L} \|V_l - V_0(l)\|_2 \leq \beta.
\]

Finally, let \( F_{\beta, \nu} \) denote the set of neural networks described above. According to Proposition 5 in Appendix.D.3, where the covering number in Eq.(19) and the Gaussian complexity in Eq.(17) are bounded, we have the final result for the generalization bound of \( F_{\beta, \nu} \). See Appendix.D.3 for the details of the proof.

Theorem 7 (Generalization bound of Convolutional Neural Networks). Given the set of convolutional neural networks \( F_{\beta, \nu} \), if the loss function \( L \) satisfies Asm.2, then for any \( \delta \in (0, 1) \), with a probability of at least \( 1 - \delta \), we have

\[
\Phi(S, \delta) \leq O\left( \frac{W L_1 \sqrt{C} \log(nC B_{\beta, \nu, \chi})}{\sqrt{n}} \right) + O\left( \frac{W L_2 \log(nB_{\beta, \nu, \chi})}{\sqrt{n}} \right),
\]
where \( B_{\beta, \nu, \chi} := \chi(1 + \nu + \beta/\alpha)^{\nu} \), \( \alpha = L_c + L_f \), and \( W \) is the number of parameters in \( F_{\beta, \nu} \).

Remark 8. According to Thm.3 and the generalization bound, it seems that we can simply set \( K \) as the number of classes. However, only the ranking of the ground-truth label and the ambiguous labels is of interest, which are all top-ranked classes. In this sense, a large \( K \) will provide a biased estimation of the performance by including more uninformative loss terms when taking the average. Motivated by this, a moderate \( K \), which is exactly equal to or slightly larger than the maximum number of relevant labels, might be a better choice. Since the maximum number of relevant labels is proportional to the number of classes, we set \( K \geq K \cdot C \), where \( p \) is a universal constant. Note that \( L_1 \sim O\left( \frac{1}{nC} \right) \) according to Thm.6. In this sense, \( \Phi(S, \delta) \) becomes irrelevant to \( C \).

Computational Complexity. On one hand, the proposed method shares a similar convergence rate with prior arts since it can be optimized efficiently via common optimizers such as SGD [81]. On the other hand, the additional complexity of the proposed method comes from the top-\( K \) ranking operator. With the help of heapsort, the average complexity of each sample is in order of \( O(C \log K) \). Then, the total complexity is in order of \( O(nC \log K + nK) \), where \( n \) is the number of samples. As analyzed in the revised remark for Thm.9, we prefer choosing a moderate \( K \). Thus, compared with top-\( k \) optimization, whose complexity is in order of \( O(nC + n) \), the complexity of the proposed method is still moderate.

7 Experiment

In this section, we analyze the empirical results on four benchmark datasets to validate the proposed AUTKC optimization framework as well as the theoretical results.

7.1 Datasets

We first describe the datasets used in the following discussion:

- Cifar-10 [82]: This dataset\(^1\) consists of 60,000 32 \times 32 color images in 10 classes, with 6,000 images per class. There are 50,000 and 10,000 images in the training set and the test set, respectively. At the first glance, Cifar-10 is not suitable for the context of label ambiguity since all the classes are completely mutually exclusive. For example, the class Automobile includes sedans, SUVs, things of that sort while the class Truck includes only big trucks. However, we have shown that AUTKC is a better metric than TOP-k for any \( k < K \) in Sec.3.4. We adopt this dataset to justify this argument even when no label ambiguity involves.

- Cifar-100 [82]: This dataset\(^1\) is similar to the Cifar-10, except it has 100 classes each containing 600 images. There are 500 training images and 100 testing images per class. As the number of classes increases, label ambiguity inevitably happens. For example, the class Sea tends to overlap with the class Dolphin and Seal; the class Cloud is widely distributed in outdoor classes such as House, Mountain, Plain and Skyscraper.

7.2 Competitors

Specifically, we hope to validate three arguments: (1) TOP-k is not an ideal optimization objective, as presented in Sec. 3.2; (2) AUTKC is a better metric than TOP-k since the models induced by AUTKC optimization consistently achieve better performance; (3) The hinge loss is inconsistent in AUTKC optimization. Motivated by this, the competitors include:

- Traditional loss functions for multiclass classification, such as cross-entropy loss (CE) and multi-class hinge loss (Hinge).
- Prior arts of TOP-k optimization, whose details are presented in Tab. 2. Note that TCE and L1 are induced by CE and Hinge via simple top-k truncation, respectively. Meanwhile, L2, L3 and L4 are hinge-like top-k loss functions consistent with TOP-k when no ties exist in the scores. L5 is the recent consistent variant even when ties exist. The value of k is searched in \{3, 5\} on Cifar-10 and in \{3, 5, 10\} on the other datasets.
- Our AUTKC optimization framework implemented with the hinge surrogate loss (AUTKC-Hinge):

\[
L_K^{hinge}(s, y) = \frac{1}{K} \sum_{k=1}^{K+1} \ell_{hinge}(s_y - s[k]),
\]

where \(K\) is searched in the same space as \(k\).

- Our AUTKC optimization framework implemented with the square surrogate loss (AUTKC-Sq), the exponential surrogate loss (AUTKC-Exp) and the logit surrogate loss (AUTKC-Logit), that is, \(L_{exp}^{CE}, L_{logit}^{CE}\) and \(L_{K}^{exp}\) mentioned in Corollary 2. The search space of \(K\) is same as that of \(k\), and the corresponding ablation study is shown in Sec.7.4.2.

7.3 Implementation details

Infrastructure. We carry out all the experiments on an ubuntu 16.04 server equipped with Intel(R) Xeon(R) Silver 4110 CPU and an Nvidia(R) TITAN RTX GPU. The version of CUDA is 10.2 with GPU driver version 440.44. The codes are implemented via python (v-3.8.11) [84], and the main third-party packages include pytorch (v-1.9.0), numpy (v-1.20.3), scikit-learn (v-0.24.2) and torchvision (v-0.10.0).

Evaluation Metric. Given a trained score function \(f\), we evaluate its performance with AUTKC↑ (the up the better):

\[
AUTKC↑ = \frac{1}{nK} \sum_{i=1}^{n} \left\{ \left| \{k : f(x_i)_{y_i} > f(x_i)[k], k \leq K + 1\} \right| \right\}.
\]

The value of \(K\) ranges in \{3, 5\} on Cifar-10 and \{3, 5, 10\} on the other datasets. Meanwhile, we also record the performance at some points of the top-k curve:

\[
TOP-k↑ = \frac{1}{n} \sum_{i=1}^{n} 1 \left\{ f(x_i)_{y_i} > f(x_i)[k] \right\},
\]

where the value of \(k\) is in the range of \[1, 5\] on Cifar-10 and [1, 10] on the other datasets.

Backbone and Optimization Method. We utilize ResNet-18 [20] as the backbone on Cifar-10, Cifar-100 and Tiny-imagenet-200. The parameters of this neural network are initialized with the checkpoint pre-trained on ILSVRC2012 [5], and we finetune the whole model with Stochastic Gradient Descent (SGD) [81] with Nesterov momentum. Empirically, the batch size and momentum are set as 128 and 0.9, respectively; the learning rate at the first epoch is searched in \{0.01, 0.001, 0.0001\}, and decays exponentially with a ratio of 0.1 every 30 epochs; the \(\ell_2\) regularization term is searched in \{0.01, 0.001, 0.0001, 0.00001\}. On Places-365, to scale up the training period, we extract the original images to the features with 2,048 dimensions via the official pre-trained ResNet-50⁴. Then, the backbone is implemented with a three-layer fully-connected neural network, whose parameters are randomly initialized. The optimization strategy is similar to that on the other datasets, except that the batch size is 4096 and the initial learning rate is searched in \{0.1, 0.01, 0.001, 0.0001\}. The total number of epochs is set as 50 on Tiny-imagenet-200 and 90 on the other datasets. Note that Tiny-imagenet-200 has the same domain as ILSVRC2012, which helps the training process converge in fewer epochs.

Warm-Up Training. Focusing on a few classes during the early training period makes it easy to suffer from overfitting. Thus, we adopt a warm-up strategy to encourage the attention on all the classes. Specifically, the model is trained with CE at the first \(E_w\) epochs. Afterward, we

### Table 2

The prior arts of TOP-k optimization, where \(1_y := 1 - \delta_y\). In these methods, TCE is a truncated variant of the traditional cross-entropy loss; and the others are induced by the multiclass hinge loss.

<table>
<thead>
<tr>
<th>Name</th>
<th>Formulation</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>[1 + (s_{y})_{[k]} - s_y + ]</td>
<td>[7], [83]</td>
</tr>
<tr>
<td>L2</td>
<td>[1/k \sum_{j=1}^{k} (s + 1_y)_{[j]} - s_y + ]</td>
<td>[4], [7], [8]</td>
</tr>
<tr>
<td>L3</td>
<td>[1/k \sum_{j=1}^{k} (s + 1_y)_{[j]} - s_y + ]</td>
<td>[4], [7], [8]</td>
</tr>
<tr>
<td>L4</td>
<td>[1/k \sum_{j=1}^{k} (1 + (s_{y})_{[j]} - s_y + ]</td>
<td>[9]</td>
</tr>
<tr>
<td>L5</td>
<td>[1 + s_{[k+1]} - s_y + ]</td>
<td>[9]</td>
</tr>
<tr>
<td>TCE</td>
<td>[log(1 + \sum_{j=1}^{k} \exp(s_{y} - s_{[j]}) + ]</td>
<td>[4], [8]</td>
</tr>
</tbody>
</table>

TABLE 3
The empirical results of AUTKC on the four datasets. The Best and the runner-up methods on each metric are marked with red and blue, respectively. The best method of the competitors is marked with underline.

<table>
<thead>
<tr>
<th>Type</th>
<th>Dataset</th>
<th>Cifar-10</th>
<th>Cifar-100</th>
<th>Tiny-imagenet-200</th>
<th>Places-365</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K$</td>
<td>3 5 10</td>
<td>3 5 10</td>
<td>3 5 10</td>
<td>3 5 10</td>
</tr>
<tr>
<td>Traditional</td>
<td>CE</td>
<td>92.87</td>
<td>95.24</td>
<td>71.81</td>
<td>76.70</td>
</tr>
<tr>
<td></td>
<td>Hinge</td>
<td>92.75</td>
<td>95.20</td>
<td>71.33</td>
<td>76.65</td>
</tr>
<tr>
<td>Top-k</td>
<td>L1</td>
<td>89.58</td>
<td>93.25</td>
<td>69.90</td>
<td>76.19</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>92.04</td>
<td>94.70</td>
<td>71.34</td>
<td>77.19</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>92.56</td>
<td>94.92</td>
<td>71.98</td>
<td>76.78</td>
</tr>
<tr>
<td></td>
<td>L4</td>
<td>91.64</td>
<td>94.48</td>
<td>71.47</td>
<td>77.28</td>
</tr>
<tr>
<td></td>
<td>L5</td>
<td>88.88</td>
<td>92.79</td>
<td>68.88</td>
<td>75.27</td>
</tr>
<tr>
<td></td>
<td>TCE</td>
<td>90.29</td>
<td>93.74</td>
<td>71.41</td>
<td>76.99</td>
</tr>
<tr>
<td></td>
<td>AUTKC-Hinge</td>
<td>92.93</td>
<td>95.27</td>
<td>73.29</td>
<td>78.33</td>
</tr>
<tr>
<td>AUTKC (Ours)</td>
<td>AUTKC-Sq</td>
<td>93.00</td>
<td>95.01</td>
<td>73.39</td>
<td>78.43</td>
</tr>
<tr>
<td></td>
<td>AUTKC-Exp</td>
<td>93.01</td>
<td>95.30</td>
<td>73.60</td>
<td>78.70</td>
</tr>
<tr>
<td></td>
<td>AUTKC-Logit</td>
<td>93.05</td>
<td>95.30</td>
<td>73.29</td>
<td>78.42</td>
</tr>
</tbody>
</table>

Fig. 6. The normalized top-$k$ accuracy gain with respect to CE on the four datasets. The scattered points represent the average performance of each method type, and the shadow represents the standard deviation.

7.4 Results and Analysis

7.4.1 Overall Performance

The Comparison on AUTKC. First, we record the best performance of each method on each dataset in Tab.3. From the results, we have the following observations:

- The AUTKC methods outperform all the competitors consistently on all the metrics on Cifar-100, Tiny-imagenet-200 and Places-365. This validates the effectiveness of the proposed optimization framework on the datasets where label ambiguity exists.
- A slight improvement is also attainable with AUTKC optimization methods on Cifar-10, where the number of classes is small. This result evidences our theoretical results in Sec.3.4 and Sec.5.1, that is, AUTKC is more discriminating than TOP-$k$.
- Among the four surrogate losses for AUTKC optimization, the hinge loss performs inferior to the others, which is consistent with our theoretical analysis in Sec.5.
- Top-$k$ optimization methods achieve a slight improvement w.r.t. CE on Cifar-100, Tiny-imagenet-200 and Places-365. This empirical result is consistent with the observations in [8], [9].
- On Cifar-10, the performances of top-$k$ optimization methods fail to surpass CE, which also validates the limitation of TOP-$k$ shown in Sec.3.2.

The Comparison on the Top-$k$ Curve. Then, we plot the top-$k$ curve in Fig.6 for more fine-grained comparison. Note that we summarize the results of each type of method for concise illustration, where the scattered points represent the average performance of each type of method, and the shadow represents the standard deviation. Meanwhile, the top-$k$ accuracy ranges from 50 to 90 at different $k$ points, while the performance gap between different methods is significantly smaller. This scale difference makes the performance gap almost invisible. To fix this issue, we normalize the performance gap on each dataset and each $k$ point by the following strategy:

$$\text{Normalized Top-}k\text{ Acc Gain of method } m = \begin{cases} \frac{\text{Gain}(m, k)}{G_+}, & \text{Gain}(m, k) > 0; \\ \text{Gain}(m, k), & \text{otherwise}. \end{cases}$$

where $\text{Gain}(m, k)$ is the top-$k$ accuracy gain of method $m$ with respect to CE, and

$$G_+ := \max_{m,k} \{\text{Gain}(m, k)\},$$

$$G_- := \min_{m,k} \{\text{Gain}(m, k)\}$$

are the maximum and minimum of all the gains on the dataset, respectively. According to Fig.6, we have the following observations:
Fig. 7. Case study on Places-365. For each input, we select the annotated label, ambiguous labels, and some irrelevant labels to visualize the ranking result produced by different types of methods. In each ranking list, the annotated label and the ambiguous labels are highlighted with green and blue box, respectively.

7.4.2 Sensitivity analysis

The Effect of $K$. In Fig. 8, we show the sensitivity in terms of hyperparameter $K$ on Places-365 for the AUTKC methods with $E_w = 40$. Note that the space of $K$ includes the total number of classes. We have the following observations from the results:

- AUTKC-Hinge has a different trend from the other methods. To be specific, its median performance is
Fig. 8. Sensitivity analysis of AUTKC optimization methods on $K$, where the experiments are conducted on Places-365. For each box, $K$ is fixed as the y-axis value, and the scattered points represent the performance with different hyperparameters.

Fig. 9. Sensitivity analysis of AUTKC optimization methods on $E_w$, where the experiments are conducted on Cifar-100. For each box, $E_w$ is fixed as the y-axis value, and the scattered points represent the performance with different hyperparameters.

insensitive to the value of $K$, and the worst performance becomes worse as the value of $K$ increases. This counter-intuitive phenomenon again validates Thm.5, that is, AUTKC-Hinge is not AUTKC consistent.

- The consistent methods, e.g., AUTKC-Sq, AUTKC-Exp and AUTKC-Logit, achieve the best median performance on AUTKC@5 when $K$ is set as 3 or 5. This observation is consistent with our goal to design the AUTKC objective.

- When $K$ equals 1, the consistent methods generally have the worst median performance. This is not surprising since this case does not properly handle label ambiguity.

- The consistent methods with $K = 365$ also perform competitively but are inferior to the ones with $K = 3$ or $K = 5$. This confirms the necessity to select a moderately large $K$.

The Effect of $E_w$. Furthermore, we show the sensitivity in terms of hyperparameter $E_w$ on Cifar-100 for the AUTKC methods in Fig.9. Note that we record the results only when the method succeeds to converge. From the results, one can observe that:

- On one hand, the increasing $E_w$ from 5 to 10 brings an increasing trend of the median performance. On the other hand, a larger $K$ generally leads to a decreasing trend of the performance deviation. This shows that the warm-up strategy is beneficial to the performance of the AUTKC methods.

- The AUTKC methods achieve the best median performance when $E_w$ is equal to 10 or 20. This suggests that a moderate warm-up strategy might be a better strategy for AUTKC optimization.

8 CONCLUSION

This paper proposes a new metric named AUTKC to overcome the limitations of the top-$k$ objective in the context of label ambiguity. An empirical surrogate risk minimization framework is further established for efficient AUTKC optimization. Theoretically, we show that the surrogate loss is Fisher consistent with the 0-1-loss-based AUTKC risk if it is differentiable, bounded, and strictly decreasing. Most common surrogate losses satisfy this condition when the model outputs are bounded, such as the square loss, the exponential loss, and the logit loss. However, the hinge loss, the standard surrogate loss for top-$k$ optimization, is inconsistent. Moreover, we construct the generalization bound for empirical AUTKC optimization, which is insensitive to the number of classes. Finally, the empirical results on four benchmark datasets justify the proposed framework and the theoretical results.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, in part by National Natural Science Foundation of China: U21B2038, 61931008, 62025604, 62132006, 6212200758 and 61976202, in part by the Fundamental Research Funds for the Central Universities, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB28000000, and in part by the National Postdoctoral Program for Innovative Talents under Grant BX2021298.


Zhiyong Yang received the M.Sc. degree in computer science and technology from University of Science and Technology Beijing (USTB) in 2017, and the Ph.D. degree from University of Chinese Academy of Sciences (UCAS) in 2021. He is currently a postdoctoral research fellow with the University of Chinese Academy of Sciences. His research interests lie in machine learning and learning theory, with special focus on AUC optimization, meta-learning/multi-task learning, and learning theory for recommender systems. He has authored or coauthored about 32 academic papers in top-tier international conferences and journals including TPAMI/ICML/NeurIPS/CVPR. He served as a Senior PC member for IJCAI 2021 and a reviewer for several top-tier journals and conferences such as TPAMI, TMLR, ICML, NeurIPS and ICLR.

Yuan He received his B.S. degree and Ph.D. degree from Tsinghua University, P.R. China. He is a Senior Staff Engineer in the Security Department of Alibaba Group, and working on artificial intelligence-based content moderation and intellectual property protection systems. Before joining Alibaba, he was a research manager at Fujitsu working on document analysis system. He has published more than 30 papers in computer vision and machine learning related conferences and journals including CVPR, ICCV, ICML, NeurIPS, AAAI and ACM MM. His research interests include computer vision, machine learning, and AI security.

Xiaochun Cao is a Professor of School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University. He received the B.E. and M.E. degrees both in computer science from Beijing University (BUAA), China, and the Ph.D. degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the university level Outstanding Dissertation Award. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a professor at Tianjin University. Before joining SYSU, he was a professor at Institute of Information Engineering, Chinese Academy of Sciences. He has authored and coauthored over 200 journal and conference papers. In 2004 and 2010, he was the recipients of the Piero Zamperoni best student paper award at the International Conference on Pattern Recognition. He is on the editorial boards of IEEE Transactions on Image Processing and IEEE Transactions on Multimedia, and was on the editorial board of IEEE Transactions on Circuits and Systems for Video Technology.

Qingming Huang is a chair professor in the University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He graduated with a Bachelor degree in Computer Science in 1988 and Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China. His research areas include multimedia computing, image processing, computer vision and pattern recognition. He has authored or coauthored more than 400 academic papers in prestigious international journals and top-level international conferences. He was the associate editor of IEEE Trans. on CSVT and Acta Automatic Sinica, and the reviewer of various international journals including IEEE Trans. on PAMI, IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, etc. He is a Fellow of IEEE and has served as general chair, program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICVR, PCM, BigMM, PSIVT, etc.
Appendix A: Metric Comparison between AUTKC and Top-k (Proof of Thm. 1)  
Appendix B: Reformulation of the Original Optimization Objective (Proof of Thm. 2)  
Appendix C: Consistency Analysis  
  C.1 The Bayes Optimal Function for AUTKC (Proof of Thm. 3)  
  C.2 Sufficient Condition for AUTKC Calibration (Proof of Thm. 4)  
  C.3 Inconsistency of Hinge Loss (Proof of Thm. 5)  
Appendix D: Generalization Analysis  
  D.1 Lipschitz Property of the Loss Functions (Proof of Thm. 6 and Corollary 2)  
    D.1.1 Lipschitz Property of the Square Loss  
    D.1.2 Lipschitz Property of the Exp Loss  
    D.1.3 Lipschitz Property of the Logit Loss  
  D.2 Sub-Gaussian Property of Gaussian Complexity (Proof of Proposition 3)  
  D.3 Generalization Bound of Convolutional Neural Networks (Proof of Thm. 7)  
    D.3.1 The Lipschitz Property of $F_{\beta,\nu}$  
    D.3.2 The Practical Generalization Bound of $F_{\beta,\nu}$
**APPENDIX A**

**METRIC COMPARISON BETWEEN AUTKC AND TOP-K (PROOF OF THM. 1)**

**Theorem 1.** For any $k < K$, AUTKC is strictly consistent and more discriminating than TOP-k.

**Proof.** Let

$$f(\pi) := \frac{1}{K} \sum_{k=1}^{K} 1[\pi(y) \leq k]$$

and

$$g(\pi) := 1[\pi(y) \leq k].$$

**Proof for consistency.** We first show AUTKC is consistent with TOP-k for any $k < K$. Given two predictions $a, b \in \mathbb{R}^C$ and a label $y$, it is clear that

$$f(\pi_a) > f(\pi_b) \iff \pi_a(y) < \pi_b(y) \text{ and } \pi_a(y) \leq K$$

and

$$g(\pi_a) > g(\pi_b) \iff \pi_a(y) \leq k \text{ and } \pi_b(y) > k.$$  

Thus,

$$R = \{(\pi_a, \pi_b) | \pi_a(y) \leq k, \pi_b(y) > k\}$$

and

$$|R| = k(C - k) > 0.$$  

Meanwhile, since

$$g(\pi_a) < g(\pi_b) \iff \pi_a(y) > k \text{ and } \pi_b(y) \leq k,$$

we have

$$S = \{(\pi_a, \pi_b) | K \geq \pi_a(y) > k \geq \pi_b(y), \pi_a(y) < \pi_b(y)\} = \emptyset.$$  

Therefore,

$$C = \frac{|R|}{|R| + |S|} = 1.$$  \hspace{1cm} (21)

**Proof for discrimination.** Next we show that AUTKC is more discriminating than TOP-k. On one hand, we have

$$g(\pi_a) = g(\pi_b) \iff \pi_a(y), \pi_b(y) \leq k \text{ or } \pi_a(y), \pi_b(y) > k.$$  

For (I),

$$P_1 = \{(\pi_a, \pi_b) | \pi_a(y) < \pi_b(y) \leq k\},$$

thus

$$|P_1| = (k - 1) + (k - 2) + \cdots + 1 = \frac{1}{2}k(k - 1).$$

For (II),

$$P_2 = \{(\pi_a, \pi_b) | k < \pi_a(y) < \pi_b(y), \pi_a(y) \leq K\},$$

thus

$$|P_2| = (C - k - 1) + \cdots + (C - K) = \frac{1}{2}(2C - k - 1)(K - k).$$

Combining (I) and (II), we have

$$|P| = |P_1| + |P_2| > 0,$$  \hspace{1cm} (22)

since $P_1 = 0$ only if $k = 1$, and $P_2 = 0$ only if $k = K$.

On the other hand, we have

$$f(\pi_a) = f(\pi_b) \iff \pi_a(y) = \pi_b(y) \text{ and } \pi_a(y), \pi_b(y) \leq K$$

or

$$\pi_a(y), \pi_b(y) > K.$$  \hspace{1cm} (III)

For (III), $Q_1 = \emptyset$ since $g(\pi_a) > g(\pi_b)$ means $\pi_a(y) \neq \pi_b(y)$. For (IV),

$$Q_2 = \{(\pi_a, \pi_b) | K < \pi_a(y) \leq k, \pi_b(y) > K\} = \emptyset.$$
Then, Combining (III) and (IV), we have
\[ |Q| = |Q_1| + |Q_2| = 0. \] (23)

With Eq.(22) and Eq.(23), we have
\[ D = \frac{|P|}{|Q|} = \infty. \] (24)

Finally, Combining Eq.(21) and Eq.(24), the proof ends.

\[ \square \]

**APPENDIX B**

**REFORMULATION OF THE ORIGINAL OPTIMIZATION OBJECTIVE (PROOF OF THM. 2)**

**Theorem 2.** When Asm.1 holds, the original optimization problem \((OP_0)\) is equivalent to the following problem:
\[
(\text{OP}_1) \min_f \mathbb{E}_{z \sim D} \left[ \frac{1}{K} \sum_{k=1}^{K+1} \ell_0 - 1 \left( s_y - s_k \right) \right].
\] (9)

**Proof.** According to Assumption 1, we have
\[
\sum_{k \leq K} \mathbf{1}[s_y \leq (s_{\setminus y})_{[k]}] = \begin{cases} 
-1 + \sum_{k \leq K} \mathbf{1}[s_y \leq s_k], & s_y > (s_{\setminus y})_{[K]}, \\
K, & \text{otherwise}
\end{cases}
\]
\[
= -1 + \mathbf{1}[s_y \leq (s_{\setminus y})_{[K]}] + \sum_{k \leq K} \mathbf{1}[s_y \leq s_k]
\]
\[
= -1 + \sum_{k \leq K+1} \mathbf{1}[s_y \leq s_k].
\]

Then the proof ends. \[ \square \]

**APPENDIX C**

**CONSISTENCY ANALYSIS**

C.1 The Bayes Optimal Function for AUTKC (Proof of Thm. 3)

**Theorem 3.** The score function \( f : \mathcal{X} \to \mathbb{R}^C \) is AUTKC Bayes optimal if and only if the score is top-\( K \) ranking-preserving with respect to the conditional risk. More concretely, we have
\[
f(x)_{y_i} > f(x)_{y_2} \text{ if } \eta(x)_{y_i} > \eta(x)_{y_2} \geq \eta(x)_{[K]}.\]

**Proof.** Since the samples are independently drawn, we define the conditional risk:
\[
\mathcal{R}_K(f, \eta(x)) = \mathbb{E}_{y \mid x \sim \eta(x)} \left[ aerr_K(y, f(x)) \right].
\] (25)

Then, we only need to show that given any sample \( x \), if \( \text{RP}_K(f^*(x), \eta(x)) \) and \( -\text{RP}_K(f(x), \eta(x)) \), then \( \mathcal{R}_K(f^*, \eta(x)) < \mathcal{R}_K(f, \eta(x)) \). For the sake of conciseness, we denote \( \eta(x) \) as \( \eta \) in the following discussion, if there exists no ambiguity. According to the definition of conditional risk, we have:
\[
\mathcal{R}_K(s, \eta) = \mathbb{E}_{y \mid x \sim \eta} \left[ \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}[\pi_*^y(y) > k] \right]
\]
\[
= \frac{1}{K} \sum_{y} \eta_y \left[ \sum_{k=1}^{K} \mathbf{1}[\pi_*^y(y) > k] \right]
\]
\[
= \frac{1}{K} \sum_{y} \eta_y \cdot \min \{\pi_*^y(y) - 1, K\}.
\] (26)

Without loss of generality, we assume that \( \eta_1 > \eta_2 > \cdots > \eta_C \). Then, for any \( s^* \) such that \( \text{RP}_K(s^*, \eta) \), we have
\[
\mathcal{R}_K(s^*, \eta) = \frac{1}{K} \eta^T \mathbf{P}_{s^*}.
\]
where
\[ P_{s^*} := [0, 1, \cdots, K - 1, \underbrace{K, \cdots, K}_{(C-K) \text{ elements}}] \in \mathbb{R}^C \]

Then, for any score \( s \) with \( -\text{RP}_K(s, \eta) \), we have
\[ \mathcal{R}_K(s, \eta) = \frac{1}{K} \eta^T P_s, \]
where \( P_s \) is a permutation of \( P_{s^*} \), according to \( \pi_s(y) \). Notice that \( P_s \) can be represented as the combination of a series of swapping operations on \( P_{s^*} \), which is defined as follows:

**Definition 12 (Top-\( K \) ranking-swapped).** For any \( y_1, y_2 \in \mathcal{Y} \) such that \( \eta_{y_1} > \max\{\eta_{y_2}, \eta_{K}\} \), given \( s \) such that \( \pi_{s_1}(y_1) < \pi_{s_2}(y_2) \), if \( s_2 \) satisfies:

- \( \pi_{s_2}(y_1) = \pi_{s_1}(y_2) \) and \( \pi_{s_2}(y_2) = \pi_{s_1}(y_1) \);
- \( \pi_{s_2}(y) = \pi_{s_1}(y) \) for any \( y \notin \{y_1, y_2\} \),

then we say \( s_2 \) is top-\( K \) ranking-swapped with respect to \( s_1 \), denoted as \( \text{RS}_K(s_2, s_1) \).

Next, we only need to show \( \mathcal{R}_K(s_1, \eta) \leq \mathcal{R}_K(s_2, \eta) \) if \( \text{RS}_K(s_2, s_1) \), with equality if and only if \( \text{RP}_K(s_2, s_1) \). In other words, any swapping operation for such \( y_1 \) and \( y_2 \) will not lead to a smaller conditional risk than \( \mathcal{R}_K(s_1, \eta) \). Since \( s \) is the combination of some RS with respect to \( s^* \), we could conclude that for any \( s \) satisfying \( -\text{RP}_K(s, \eta) \), \( \mathcal{R}_K(s^*, \eta) < \mathcal{R}_K(s, \eta) \) holds.

To show \( \mathcal{R}_K(s_1, \eta) \leq \mathcal{R}_K(s_2, \eta) \) if \( \text{RS}_K(s_2, s_1) \), we consider the following three situations:

- **S1:** If \( \pi_{s_1}(y_1) < \pi_{s_1}(y_2) \leq K \), we have
  \[
  K \cdot [\mathcal{R}_K(s_1, \eta) - \mathcal{R}_K(s_2, \eta)] \\
  = [\eta_{y_1}(\pi_{s_1}(y_1) - 1) + \eta_{y_2}(\pi_{s_1}(y_2) - 1)] \\
  - [\eta_{y_1}\min\{\pi_{s_2}(y_1) - 1, K\} + \eta_{y_2}\min\{\pi_{s_2}(y_1) - 1, K\}] \\
  = (\eta_{y_1} - \eta_{y_2})[\pi_{s_1}(y_1) - \pi_{s_1}(y_2)] \\
  < 0. 
  \]

- **S2:** If \( \pi_{s_1}(y_1) \leq K < \pi_{s_1}(y_2) \), we have
  \[
  K \cdot [\mathcal{R}_K(s_1, \eta) - \mathcal{R}_K(s_2, \eta)] \\
  = [\eta_{y_1}(\pi_{s_1}(y_1) - 1) + \eta_{y_2}K] - [\eta_{y_1}\min\{\pi_{s_2}(y_1) - 1, K\}] \\
  = (\eta_{y_1} - \eta_{y_2})[\pi_{s_1}(y_1) - K] \\
  < 0. 
  \]

- **S3:** If \( K < \pi_{s_1}(y_1) < \pi_{s_1}(y_2) \), we have
  \[
  K \cdot [\mathcal{R}_K(s_1, \eta) - \mathcal{R}_K(s_2, \eta)] \\
  = [\eta_{y_1}K + \eta_{y_2}K] - [\eta_{y_1}\min\{\pi_{s_2}(y_1) - 1, K\}] \\
  = 0. 
  \]

Note that in S3, \( \text{RP}_K(s_2, s_1) \) holds, then the proof ends with the transitivity of \( \text{RP}_K \).

### C.2 Sufficient Condition for AUTKC Calibration (Proof of Thm. 4)

**Theorem 4.** The surrogate loss \( \ell(t) \) is AUTKC consistent if it is bounded, differentiable and strictly decreasing.

**Proof.** We first define the conditional risk and the optimal conditional risk of the surrogate loss:
\[
\mathcal{R}_K^t(s, \eta) = \mathbb{E}_{y|z \sim \eta} \left[ \frac{1}{K} \sum_{k=1}^{K+1} \ell(s_y - s_{[k]}) \right]. 
\]
\[
\mathcal{R}_K^{t*}(\eta) := \inf_{s \in \mathbb{R}^C} \mathcal{R}_K^t(s, \eta). 
\]

Then, we prove the theorem by the following steps.

**Claim 1.** If \( s^* \in \arg \inf_s \mathcal{R}_K^t(s, \eta) \), then \( \text{RP}(s^*, \eta) \).
We show this claim by showing that if \(-\text{RP}_K(s, \eta)\), then \(\mathcal{R}^\ell_K(s, \eta) > \mathcal{R}^{\ell^*}_K(\eta)\). Note that two cases will lead to \(-\text{RP}_K(s, \eta)\):

**Case 1:** If ties exist in \(\{s[1], \ldots, s[K], s[K+1]\}\), the chain is obtained by a contradiction. Given \(s\) such that \(\pi_s(y_1), \pi_s(y_2) \leq K + 1\) and \(s_{y_1} = s_{y_2}\), we assume that \(\mathcal{R}^\ell_K(s, \eta) = \mathcal{R}^{\ell^*}_K(\eta)\). According to the first-order condition, we have:

\[
\frac{\partial}{\partial s_y} \mathcal{R}^\ell_K(s, \eta) = 0, y = y_1, y_2.
\]

That is,

\[
\eta_{y_1} \sum_{k=1,|k| \neq y_1}^{K+1} \ell'(s_{y_1} - s_{[k]}) = \sum_{y \neq y_1} \eta_y \ell'(s_y - s_{y_1});
\]

\[
\eta_{y_2} \sum_{k=1,|k| \neq y_2}^{K+1} \ell'(s_{y_2} - s_{[k]}) = \sum_{y \neq y_2} \eta_y \ell'(s_y - s_{y_2}),
\]

where \([k] \neq y_1\) means that the calculation of the derivative will be skipped when \(\pi_s(y_1) = k\). Since \(s_{y_1} = s_{y_2}\), we have

\[
(I) = \eta_{y_1} \ell'(0) + \eta_{y_1} \sum_{k=1,|k| \neq \{y_1, y_2\}}^{K+1} \ell'(s_{y_1} - s_{[k]});
\]

\[
(II) = \eta_{y_2} \ell'(0) + \sum_{y \neq \{y_1, y_2\}} \eta_y \ell'(s_y - s_{y_1});
\]

\[
(III) = \eta_{y_2} \ell'(0) + \eta_{y_2} \sum_{k=1,|k| \neq \{y_1, y_2\}}^{K+1} \ell'(s_{y_1} - s_{[k]});
\]

\[
(IV) = \eta_{y_1} \ell'(0) + \sum_{y \neq \{y_1, y_2\}} \eta_y \ell'(s_y - s_{y_1});
\]

Then, since \((I) - (III) = (II) - (IV)\), we have

\[
(\eta_{y_1} - \eta_{y_2}) \left[ \ell'(0) + \sum_{y \neq \{y_1, y_2\}} \eta_y \ell'(s_y - s_{y_1}) \right] = (\eta_{y_2} - \eta_{y_1}) \ell'(0),
\]

which leads to a contradiction since \(\eta_{y_2} \neq \eta_{y_1}\) and \(\ell'(t) < 0\).

**Case 2:** If no ties in \(\{s[1], \ldots, s[K], s[K+1]\}\), for any \(s\) with \(-\text{RP}_K(s, \eta)\), we can construct an \(s^*\) with \(\text{RP}_K(s^*, \eta)\) by a series of top-\(K\) ranking-resuming operation, that is, the inverse operation of the top-\(K\) ranking-swapping operation. Since \(\mathcal{R}^\ell_K(s^*, \eta) \geq \mathcal{R}^{\ell^*}_K(\eta)\), we only need to show any top-\(K\) ranking-resuming operation will lead to a smaller conditional risk. Specifically, for any \(s_1, s_2\) satisfying

- \(\eta_{y_1} > \eta_{y_2}\) and \(s_{1,y_1} > s_{1,y_2}\);
- \(s_{2,y_1} = s_{1,y_2}\) and \(s_{2,y_2} = s_{1,y_1}\);
- \(s_{2,y} = s_{1,y}\) for any \(y \notin \{y_1, y_2\}\).

we need to show \(\mathcal{R}^\ell_K(s_1, \eta) < \mathcal{R}^\ell_K(s_2, \eta)\). On top of these properties, we have \(s_{1,[k]} = s_{2,[k]}\) for any \(k \in [C]\). Then,

\[
K \cdot \left[ \mathcal{R}^\ell_K(s_1, \eta) - \mathcal{R}^\ell_K(s_2, \eta) \right]
\]

\[
= \left[ \eta_{y_1} \sum_{k=1}^{K+1} \ell(s_{1,y_1} - s_{1,[k]}) + \eta_{y_2} \sum_{k=1}^{K+1} \ell(s_{1,y_2} - s_{1,[k]}) \right] - \left[ \eta_{y_2} \sum_{k=1}^{K+1} \ell(s_{2,y_1} - s_{2,[k]}) + \eta_{y_2} \sum_{k=1}^{K+1} \ell(s_{2,y_2} - s_{2,[k]}) \right]
\]

\[
= \left[ \eta_{y_1} \sum_{k=1}^{K+1} \ell(s_{1,y_1} - s_{1,[k]}) + \eta_{y_2} \sum_{k=1}^{K+1} \ell(s_{1,y_2} - s_{1,[k]}) \right] - \left[ \eta_{y_2} \sum_{k=1}^{K+1} \ell(s_{1,y_2} - s_{1,[k]}) + \eta_{y_2} \sum_{k=1}^{K+1} \ell(s_{1,y_1} - s_{1,[k]}) \right]
\]

\[
= (\eta_{y_1} - \eta_{y_2}) \sum_{k=1}^{K+1} \left[ \ell(s_{1,y_1} - s_{1,[k]}) - \ell(s_{1,y_2} - s_{1,[k]}) \right] < 0
\]

The inequality holds since \(s_{1,y_1} > s_{1,y_2}\), and the surrogate loss \(\ell\) is strictly decreasing.

Then, the **Claim 1** is completed by **Case 1** and **Case 2**.

**Claim 2.**

\[
\inf_{s: \neg \text{RP}(s, \eta)} \mathcal{R}^\ell_K(s, \eta) > \inf_{s: \neg \text{RP}(s, \eta)} \mathcal{R}^\ell_K(s, \eta)
\]
Theorem 5. \( R\) hinge loss is inconsistent with two steps: (1) remove the term \( \eta \) score
Claim 2
According to Proof.
Without loss of generality, we assume Given constants Lemma 2.

C.3 Inconsistency of Hinge Loss (Proof of Thm. 5)

Claim 3. For any sequence \( \{s_t\}_{t \in \mathbb{N}_+}, s_t \in \mathbb{R}^C \),

\[
R_{\ell}^f(s_t, \eta) \rightarrow \inf_{s \in \mathbb{R}^C} R_{\ell}^f(s, \eta) = R_K(s_t, \eta) \rightarrow \inf_{s \in \mathbb{R}^C} R_K(s, \eta).
\]

On top of Claim 1, we only need to prove \( R_P(s_t, \eta) \) when \( t \to \infty \). Define

\[
\delta := \inf_{s \to R_P(s_t, \eta)} R^f_{\ell}(s, \eta) - \inf_{s \in \mathbb{R}^C} R^f_{\ell}(s, \eta).
\]

According to Claim 2, we have \( \infty > \delta > 0 \). Suppose that when \( t \to \infty \), \(-R_P(s_t, \eta)\). Then, there exists a large enough \( T \)
such that

\[
R^f_{\ell}(s_t, \eta) - \inf_{s \in \mathbb{R}^C} R^f_{\ell}(s, \eta) > \delta,
\]

which is contradicts with \( R^f_{\ell}(s_t, \eta) \to \inf_{s \in \mathbb{R}^C} R^f_{\ell}(s, \eta) \). This shows that Claim 3 holds.

Claim 4. For any sequence \( \{f_t\}_{t \in \mathbb{N}_+}, f_t \in \mathcal{F} \),

\[
R_{\ell}^f(f_t) \to \inf_{f \in \mathcal{F}} R_{\ell}^f(f) = R_K(f_t) \to \inf_{f \in \mathcal{F}} R_K(f).
\]

It is clear that Claim 4 holds with Claim 3 and

\[
R_{\ell}^f(f_t) = \mathbb{E}_x \left[ R_{\ell}^f(f_t(x), \eta(x)) \right]
\]

Then, the proof of Thm.4 ends.

C.3 Inconsistency of Hinge Loss (Proof of Thm. 5)

We first present the following simple but useful lemma:

Lemma 2. Given constants \( a, b > 0 \), the function \( F(t) = b[1 + t]_+ + a[1 - t]_+ \) is strictly increasing when \( t > 1 \).

Proof. It is clear that \( F(t) = b(1 + t) \) when \( t > 1 \), then the proof ends.
On one hand,
\[(IV) - (II) = \sum_{y=K+2}^{K+1} \eta_y \sum_{k=1}^{K+1} (s_k^* - s_{K+1}) \tag{32}\]
is increasing as \(\Delta_y := s_y^* - s_k^*\) increases when \(y < k \leq K + 1\). On the other hand, according to Lem.2,
\[(III) = \sum_{y=1}^{K+1} \eta_y + \sum_{y=1}^{K+1} \sum_{k=y+1}^{K+1} (\eta_y [1 + s_k^* - s_y^*] + \eta_k [1 + s_y^* - s_k^*]) \tag{33}\]
is also increasing w.r.t. \(\Delta_{y,k}\) when \(\Delta_{y,k} > 1\). Combining Eq.(32) and Eq.(33), we conclude that if exists \(y < k \leq K + 1\) such that \(\Delta_{y,k} > 1\), then
\[\mathcal{R}_K^t(s^*, \eta) - \mathcal{R}_K^t(s, \eta) > \inf_{s^* : \text{RP}(s^*, \eta)} \left[ \mathcal{R}_K^t(s^*, \eta) - \mathcal{R}_K^t(s, \eta) \right] . \]
Thus, we next assume \(\Delta_{y,k} \leq 1\) for any \(y, k \leq K + 1\). In this case, for any \(y \leq K\) we have
\[
\frac{\partial}{\partial s_y^*} \left[ \mathcal{R}_K^t(s^*, \eta) - \mathcal{R}_K^t(s, \eta) \right] = -\eta_y + \frac{1}{K} \sum_{k=1, k \neq y}^{K+1} \eta_k \\
= -\eta_y + \frac{1}{K} (1 - \eta_y) \\
> - \frac{1}{K+1} + \frac{1}{K} (1 - \frac{1}{K+1}) = 0
\]
Note that \(s_1^* > \cdots > s_K^* > s_{K+1}^*\), which means:
\[\mathcal{R}_K^t(s^*, \eta) - \mathcal{R}_K^t(s, \eta) \geq \lim_{s_y^* \rightarrow s_{[K+1]}, y \leq K} \left[ \mathcal{R}_K^t(s^*, \eta) - \mathcal{R}_K^t(s, \eta) \right] = 0.\]
The equality holds only when \(s_y^* = s_{[K+1]}^*, y \leq K\), which contradicts with the fact \(\text{RP}(s^*, \eta)\). Thus, we have \(\mathcal{R}_K^t(s^*, \eta) > \mathcal{R}_K^t(s, \eta)\), which completes the proof.

\[\square\]

**APPENDIX D**

**GENERALIZATION ANALYSIS**

**D.1 Lipschitz Property of the Loss Functions (Proof of Thm. 6 and Corollary 2)**

**Theorem 6.** Assume that dom \(s_i \subset [0, B_i]\) for any \(i = 1, \cdots, C\), if \(\ell(t)\) is strictly decreasing and \(L_\ell\)-Lipschitz continuous in the range \([-B_s, B_s]\), Then
\[L_K(s, y) = \frac{1}{K} \sum_{k=1}^{K+1} \ell \left( s_y - s_{[k]} \right) \]
is Lipschitz continuous w.r.t. a variant of the \(\ell_2\)-norm with Lipschitz constant pair \(\left( \frac{L_\ell \sqrt{K + 1}}{K}, \frac{L_\ell (K+1)}{K} \right)\).

**Proof.** According to definition of \(L_K\), we have
\[|L_K(s, y) - L_K(s', y)| \]
\[\leq \frac{1}{K} \sum_{k=1}^{K+1} |\ell(s_y - s_{[k]}) - \ell(s'_y - s'_{[k]})| \]
\[\leq \frac{1}{K} \max_{k \leq K+1} \sum_{k=1}^{K+1} |\ell(s_y - s_k) - \ell(s'_y - s'_k)| \]
\[\leq \frac{1}{K} \max_{k \leq K+1} \sum_{k=1}^{K+1} |\ell(s_y - s_k) - \ell(s'_y - s'_k)| \]

\[\leq \frac{1}{K} \max_{k \leq K+1} \sum_{k=1}^{K+1} |\ell(s_y - s_k) - \ell(s'_y - s'_k)| \]

\[\leq \frac{1}{K} \max_{k \leq K+1} \sum_{k=1}^{K+1} |\ell(s_y - s_k) - \ell(s'_y - s'_k)| \]

\[\leq \frac{1}{K} \max_{k \leq K+1} \sum_{k=1}^{K+1} |\ell(s_y - s_k) - \ell(s'_y - s'_k)| \]

\[\leq \frac{1}{K} \max_{k \leq K+1} \sum_{k=1}^{K+1} |\ell(s_y - s_k) - \ell(s'_y - s'_k)| \]
The second equality holds since $\ell$ is strictly decreasing; the last equality holds since
\[
K \cdot \max_{k \leq K} t_k, \forall t \in \mathbb{R}^C;
\]
The first inequality holds since
\[
|\max \{a_1, \ldots, a_K\} - \max \{b_1, \ldots, b_K\}| \leq \max \{|a_1 - b_1|, \ldots, |a_K - b_K|\}, \quad \forall a, b \in \mathbb{R}^K
\] (34)
Since $\ell$ is $L_\ell$-Lipschitz continuous, the last term is bounded by
\[
\frac{L_\ell}{K} \max_{k \leq K+1} \sum_{k=1}^{K+1} |(s_k - s'_k) - (s_y - s'_y)|
\]
\[
\leq \frac{L_\ell}{K} \max_{k \leq K+1} \sum_{k=1}^{K+1} |(s_k - s'_k)| + \frac{L_\ell(K + 1)}{K} |s_y - s'_y|
\]
\[
\leq \frac{L_\ell \sqrt{K + 1}}{K} \max_{k \leq K+1} \left[ \sum_{k=1}^{K+1} (s_k - s'_k)^2 \right]^{\frac{1}{2}} + \frac{L_\ell(K + 1)}{K} |s_y - s'_y|
\]
\[
\leq \frac{L_\ell \sqrt{K + 1}}{K} \left[ \sum_{k=1}^{C} (s_k - s'_k)^2 \right]^{\frac{1}{2}} + \frac{L_\ell(K + 1)}{K} |s_y - s'_y|
\]

Then, the proof ends according to Def.2.

To obtain Corollary 2, the following lemma is necessary.

**Lemma 3** (Lipschitz Constant for Softmax Function [85]). Given $t \in \mathbb{R}^C$, the softmax function
\[
\text{soft}_{i}(t) = \frac{t_i}{\sum_{j=1}^{C} \exp(t_j)}
\]
is $\sqrt{2} -$Lipschitz continuous w.r.t. vector $\ell_2$ norm.

**Corollary 2.** Let $\overline{s}$ denote the normalized score $\text{soft}(s)$, where $\text{soft}$ is the softmax function. Then, the following statements hold:

- The square loss
\[
L_{\text{sq}}^K(s, y) = \frac{1}{K} \sum_{k=1}^{K+1} \ell_{\text{sq}}(\overline{s}_y - \overline{s}_k)
\]
is Lipschitz continuous w.r.t. a variant of the $\ell_2$-norm with Lipschitz constant pair
\[
\left( \frac{2\sqrt{2}(K + 1)}{K}, \frac{2\sqrt{2}(K + 1)}{K} \right).
\]

- The exponential loss
\[
L_{\text{exp}}^K(s, y) = \frac{1}{K} \sum_{k=1}^{K+1} \ell_{\text{exp}}(\overline{s}_y - \overline{s}_k)
\]
is Lipschitz continuous w.r.t. a variant of the $\ell_2$-norm with Lipschitz constant pair
\[
\left( \frac{e\sqrt{2}(K + 1)}{2K}, \frac{e\sqrt{2}(K + 1)}{2K} \right).
\]

- The logit loss
\[
L_{\text{logit}}^K(s, y) = \frac{1}{K} \sum_{k=1}^{K+1} \ell_{\text{logit}}(\overline{s}_y - \overline{s}_k)
\]
is Lipschitz continuous w.r.t. a variant of the $\ell_2$-norm with Lipschitz constant pair
\[
\left( \frac{\sqrt{2}K + 1}{2eK \ln 2}, \frac{\sqrt{2}(K + 1)}{2eK \ln 2} \right).
\]

**D.1.1 Lipschitz Property of the Square Loss**

**Proof.** It is clear that $\ell_{\text{sq}}(t)$ is strictly decreasing in the range $[-1, 1]$, and $|\ell'_{\text{sq}}(t)| \leq 4$. Then, the proof ends by Lem.6 and Lem. 3.
D.1.2 Lipschitz Property of the Exp Loss
Proof. It is clear that \( \ell_{\text{exp}}(t) \) is strictly decreasing in the range \([-1, 1]\), and \( |\ell'_{\text{exp}}(t)| \leq e \). Then, the proof ends by Thm. 6 and Lem. 3.

D.1.3 Lipschitz Property of the Logit Loss
Proof. It is clear that \( \ell_{\text{logit}}(t) \) is strictly decreasing in the range \([-1, 1]\), and \( |\ell'_{\text{logit}}(t)| \leq \frac{1}{1+e^2} \). Then, the proof ends by Lem. 6 and Lem. 3.

D.2 Sub-Gaussian Property of Gaussian Complexity (Proof of Proposition 3)
Proposition 3 (Sub-Gaussian Property of Gaussian Complexity). For any \( f \in \mathcal{F} \), let
\[
T_f(g) = \sum_{i=1}^{n} \sum_{j=1}^{C} g_{i,j} s_j^{(i)},
\]
Then, \( Z := \frac{1}{\sqrt{nC}}(T_f(g) - T_f'(g)) \) is a sub-Gaussian process with metric
\[
d_{\infty,S}(s, s') = \max_{z \in S} |s_y - s'_y|,
\]
where \( s_j^{(i)} = f(x^{(i)})(j) \).
Proof. It is clearly \( Z \) is a Gaussian distribution with \( E[Z] = 0 \). Meanwhile, we have
\[
\text{Var}[Z] = \text{Var}\left[ \frac{1}{\sqrt{nC}} \sum_{i=1}^{n} \sum_{j=1}^{C} (s_{i,j} - s'_{i,j}) g_{i,j} \right]
\leq \frac{1}{nC} \sum_{i=1}^{n} \sum_{j=1}^{C} (s_{i,j} - s'_{i,j})^2 \text{Var}[g_{i,j}]
\leq \max_{i,j} (s_{i,j} - s'_{i,j})^2
= d_{\infty, S}^2(s, s'),
\]
where the second equality holds since if \( g_{i,j}, i \in [n], j \in [C] \) are independent from each other, we have
\[
\text{Var}[\sum_{i} g_{i,t_i}] = \sum_{i} \text{Var}[g_{i,i}];
\]
the third equality is due to \( \text{Var}[g_{i,j}] = 1 \). Further, the moment-generating function of \( Z \) is
\[
E[e^{\lambda Z}] = E[e^{\lambda(Z - E[Z])}] \leq e^{\lambda^2 \text{Var}[Z]} \leq e^{\lambda^2 d_{\infty, S}^2(s, s')}.
\]
According to Def. 11, we could conclude that \( Z \) is a sub-Gaussian process with metric being \( d_{\infty, S}(s, s') \).

D.3 Generalization Bound of Convolutional Neural Networks (Proof of Thm. 7)
According to Proposition 2 and Proposition 4, to obtain the final generalization bound, our task is to bound the covering number of \( \mathcal{F}_{\beta, \nu} \). To this end, we first present the Lipschitz property of \( \mathcal{F}_{\beta, \nu} \) in Corollary 3. Then, the practical generalization bound for convolutional neural networks is obtained by Lem. 5 and Proposition 5.

D.3.1 The Lipschitz Property of \( \mathcal{F}_{\beta, \nu} \)
Definition 13. We say a function set \( \mathcal{F} \) is \((B, d, \cdot)\)-Lipschitz parameterized if there is a norm \( \| \cdot \| \) in \( \mathbb{R}^d \) and a mapping \( \phi \) from the unit ball \( \mathcal{X} \) such that: for any \( \mathbf{x} \in \mathcal{X} \), and \( \theta \) and \( \theta' \) with \( \| \theta \| \leq 1 \) and \( \| \theta' \| \leq 1 \), the following inequality holds:
\[
|\phi(\theta)(\mathbf{x}) - \phi(\theta')(\mathbf{x})| \leq B\|\theta - \theta'\|.
\]
Lemma 4 ([79]). \( \mathcal{F}_{\beta, \nu} \) is \((B_{\beta, \nu, X}, \nu')\)-Lipschitz parameterized, where \( B_{\beta, \nu, X} := (1 + \beta / L_a)^L \), \( L_a = L_c + L_f \), and \( \nu' \) is the number of parameters in \( \mathcal{F}_{\beta, \nu} \).

Corollary 3. \( \tilde{\mathcal{F}}_{\beta, \nu} \) is also \((B_{\beta, \nu, X}, \nu')\)-Lipschitz parameterized.
Proof. According to Lem. 4, there is a norm \( \| \cdot \| \) in \( \mathbb{R}^d \) and a mapping \( \phi \) from the unit ball \( \mathcal{X} \) such that: for any \( \mathbf{x} \in \mathcal{X} \), and \( \theta \) and \( \theta' \) with \( \| \theta \| \leq 1 \) and \( \| \theta' \| \leq 1 \), the following inequality holds:
\[
|\phi(\theta)(\mathbf{x}) - \phi(\theta')(\mathbf{x})| \leq B_{\beta, \nu, X}\|\theta - \theta'\|.
\]
Let \( \tilde{\phi} = e_y \circ \phi \), and we have
\[
|\tilde{\phi}(\theta)(z) - \tilde{\phi}(\theta')(z)| = |e_y \phi(\theta)(x) - e_y \phi(\theta')(x)| \\
\leq |e_y| \cdot |\phi(\theta)(x) - \phi(\theta')(x)| \\
\leq B_{\beta, \nu, \chi} |\theta - \theta'|
\]

Then the proof ends.

\[\square\]

### D.3.2 The Practical Generalization Bound of \( \mathcal{F}_{\beta, \nu} \)

The following lemma bridges \((B, d)\)-Lipschitz parameterized and covering number, and Proposition 5 further provides the upper bound of the Gaussian complexity in Eq.(17).

#### Lemma 5 (Upper bound of Covering Number [78]).

If the score function set \( \mathcal{F} \) is \((B, d)\)-Lipschitz parameterized, then the metric entropy defined on the metric space \((\mathcal{F}, d_{\infty, S})\) is bounded:

\[
\log \mathcal{C}(\mathcal{F}, \epsilon, d_{\infty, S}) \leq d \log \frac{3B}{\epsilon} \quad (35)
\]

#### Proposition 5.

Given the score function \( f \in \mathcal{F} \), if \( \tilde{\mathcal{F}} \) is \((B, d)\)-Lipschitz parameterized, then we have:

\[
\Phi_S(\tilde{\mathcal{F}}) \lesssim O \left( \frac{d \log BnC}{\sqrt{nC}} \right), \quad (36)
\]

and

\[
\Phi_{\tilde{S}}(\tilde{\mathcal{F}}) \lesssim O \left( \frac{d \log Bn}{\sqrt{n}} \right), \quad (37)
\]

where \( O \) is the complexity.

#### Proof.

Applying Lem. 5 to Eq. (19), we have

\[
\Phi_S(\tilde{\mathcal{F}}) \leq \inf_{\alpha \geq 0} \left( C_1 \alpha + \frac{C_2}{\sqrt{nC}} \int_0^1 d \log \frac{3B}{\epsilon} d\epsilon \right) \\
\leq \inf_{\alpha \geq 0} \left( C_1 \alpha + \frac{C_2}{\sqrt{nC}} (1 - \alpha) d \log \frac{3B}{\epsilon} \alpha \right)
\]

The last equation holds since \( d \log \frac{3B}{\epsilon} \) is non-increasing w.r.t. \( \epsilon \). By setting \( \alpha = \frac{1}{\sqrt{nC}} \), we have

\[
\Phi_S(\tilde{\mathcal{F}}) \leq \frac{1}{\sqrt{nC}} \left[ C_1 + C_2 d \left( 1 - \frac{1}{\sqrt{nC}} \right) \log 3B \sqrt{nC} \right]. \quad (38)
\]

Similarly, we have

\[
\Phi_{\tilde{S}}(\tilde{\mathcal{F}}) \leq \frac{1}{\sqrt{n}} \left[ C_1 + C_2 d \left( 1 - \frac{1}{\sqrt{n}} \right) \log 3B \sqrt{n} \right]. \quad (39)
\]

Then the proof ends.

\[\square\]

Finally, the practical generalization bound of the convolutional neural networks we discuss is clear:

#### Theorem 7 (Generalization bound of Convolutional Neural Networks).

Given the set of convolutional neural networks \( \mathcal{F}_{\beta, \nu} \), if the loss function \( L \) satisfies Asm.2, then for any \( \delta \in (0, 1) \), with a probability of at least \( 1 - \delta \), we have

\[
\Phi(S, \delta) \lesssim O \left( \frac{WL_1 \sqrt{C} \log (nCB_{\beta, \nu, \chi})}{\sqrt{n}} \right) \\
+ O \left( \frac{WL_2 \log (nB_{\beta, \nu, \chi})}{\sqrt{n}} \right),
\]

where \( B_{\beta, \nu, \chi} := \chi \beta (1 + \nu + \beta/L_a)^{L_a} \), \( L_a = L_c + L_f \), and \( W \) is the number of parameters in \( \mathcal{F}_{\beta, \nu} \).

#### Proof.

The proof ends by Proposition 2, Proposition 5 and Corollary 3.

\[\square\]