Exploring the Algorithm-Dependent Generalization of AUPRC Optimization with List Stability

Peisong Wen[1,2] **Qianqian Xu**[1*] **Zhiyong Yang**[2]
Yuan He[3] **Qingming Huang**[1,2,4,5*]
[1] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS
[2] School of Computer Science and Tech., University of Chinese Academy of Sciences
[3] Alibaba Group
[4] BDKM, University of Chinese Academy of Sciences
[5] Peng Cheng Laboratory
{wenpeisong20z,xuqianqian}@ict.ac.cn
{yangzhiyong21,qmhuang}@ucas.ac.cn heyuan.hy@alibaba-inc.com

Abstract

Stochastic optimization of the Area Under the Precision-Recall Curve (AUPRC) is a crucial problem for machine learning. Although various algorithms have been extensively studied for AUPRC optimization, the generalization is only guaranteed in the multi-query case. In this work, we present the first trial in the single-query generalization of stochastic AUPRC optimization. For sharper generalization bounds, we focus on algorithm-dependent generalization. There are both algorithmic and theoretical obstacles to our destination. From an algorithmic perspective, we notice that the majority of existing stochastic estimators are biased only when the sampling strategy is biased, and is leave-one-out unstable due to the non-decomposability. To address these issues, we propose a sampling-rate-invariant unbiased stochastic estimator with superior stability. On top of this, the AUPRC optimization is formulated as a composition optimization problem, and a stochastic algorithm is proposed to solve this problem. From a theoretical perspective, standard techniques of the algorithm-dependent generalization analysis cannot be directly applied to such a listwise compositional optimization problem. To fill this gap, we extend the model stability from instancewise losses to listwise losses and bridge the corresponding generalization and stability. Additionally, we construct state transition matrices to describe the recurrence of the stability, and simplify calculations by matrix spectrum. Practically, experimental results on three image retrieval datasets on speak to the effectiveness and soundness of our framework.

1 Introduction

Area Under the Precision-Recall Curve (AUPRC) is a widely used metric in the machine learning community, especially in learning to rank, which effectively measures the trade-off between precision and recall of a ranking model. Compared with threshold-specified metrics like accuracy and recall@k, AUPRC reflects a more comprehensive performance by capturing all possible thresholds. In addition, literature has shown that AUPRC is insensitive toward data distributions [20], making it adaptable to largely skewed data. Benefiting from these appealing properties, AUPRC has become one of the standard metrics in various applications, *e.g.*, retrieval [55, 58, 22, 41], object detection [45, 49, 15], medical diagnosis [50, 35], and recommendation systems [16, 73, 1, 64, 2].

Over the past decades, the importance of AUPRC has prompted extensive researches on direct AUPRC optimization. Early work focuses on full-batch optimization [45, 44, 26]. However, in the

*Corresponding authors.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

era of deep learning, the rapidly growing scale of models and data makes these full-batch algorithms infeasible. Therefore, in recent years, it has raised an increasing favor of the stochastic AUPRC optimization [9, 12, 31, 46]. Since AUPRC optimization is a stochastic dependent compositional optimization problem, general convergence rates are infeasible for AUPRC optimization. To fill this gap, [54, 70, 69] provide AUPRC optimization algorithms with provable convergence. See Appendix A for more on related work.

Despite the promoting performance of these methods in various scenarios, the generalization of AUPRC optimization algorithms is still an open problem. Some studies [17, 63] provide provable generalization for AUPRC optimization in information retrieval. In this scene, a dataset consists of multiple queries, where each query corresponds to a set of positive and negative samples. However, these results require sufficient queries to ensure small generalization errors, but leave the single-query case alone, *i.e.*, *whether the generalization error tends to zero with the length of a single-query increasing is still unclear*. This limits the adaptation scope of these methods. To fill this gap, in this paper we aim to ***design a stochastic optimization framework for AUPRC with a provable algorithm-dependent generalization performance in the single-query case.***

The target is challenging in three aspects: **(a)** Most AUPRC stochastic estimators are biased with a biased sampling rate. Moreover, due to the non-decomposability, outputs of existing algorithms might change a lot with slight changes in the training data, which is called *leave-one-out unstable* in this paper. Such an unstability is harmful to the generalization. **(b)** The standard framework to analyze the algorithm-dependent generalization requires the objective function to be expressed as a sum of instancewise terms, while AUPRC involves a listwise loss. **(c)** The stochastic optimization of AUPRC is a two-level compositional optimization problem, which brings more complicated proofs of the stability.

In search of a solution to **(a)**, we propose a sampling-rate-invariant asymptotically unbiased stochastic estimator based on a reformulation of AUPRC. Notably, to ensure the stability [28, 39, 37, 38] of the estimator, the objective is formulated as a two-level compositional problem by introducing an auxiliary vector for the ranking estimation. Error analysis further supports the feasibility of our method, and inspires us to add a semi-variance regularization term. To solve this problem, we propose an algorithm with provable convergence that combines stochastic gradient descent (SGD), linear interpolation and exponential moving average.

Facing challenge **(b)**, we extend instancewise model stability to listwise model stability, and correspondingly put forward the generalization via stability of listwise problems. On top of this, we bridge the generalization of AUPRC and the stability of the proposed optimization algorithm.

As for challenge **(c)**, since the variables to be optimized are typically updated alternately in the compositional optimization problem, we propose state transition matrices of these variables, and simplify the calculations of the stability with matrix spectrum.

In a nutshell, the main contributions of this paper are summarized as follows:

- Algorithmically, a stochastic learning algorithm is proposed for AUPRC optimization. The core of the proposed algorithm is a stochastic estimator which is sampling-rate-invariant asymptotically unbiased.
- Theoreticallly, we present the first trial on the algorithm-dependent generalization of stochastic AUPRC optimization. To the best of our knowledge, it is also the first work to analyze the stability of stochastic compositional optimization problems.
- Technically, we extend the concept of the stability and generalization guarantee to listwise non-convex losses. Then we simplify the stability analysis of compositional objective by matrix spectrum. These techniques might be instructive for other complicated metrics.

2 Problem Formulation

2.1 Preliminaries on AUPRC

Notations. Consider a set of N examples $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ independently drawn from a sample space $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y} = \{-1, 1\}$ is the label space. For sake of the presentation, denote the set of positive examples of $\mathcal{S}$ as $\mathcal{S}^+ = \{\boldsymbol{x}_i^+\}_{i=1}^{N^+}$, and similarly the set of

negative examples is denoted as $\mathcal{S}^- = \{\boldsymbol{x}_i^-\}_{i=1}^{N^-}$, where $N^+ = |\mathcal{S}^+|, N^- = |\mathcal{S}^-|$. With a slight abuse of notation, we also denote $\mathcal{S} = \mathcal{S}^+ \cup \mathcal{S}^-$ if there is no ambiguity. Generally, we assume that the dataset is sufficiently large, such that $N^+/(N^+ + N^-) = \mathbb{P}(y = 1) := \pi$. Our target is to learn a score function $h_{\boldsymbol{w}} : \mathcal{X} \mapsto \mathbb{R}$ with parameters $\boldsymbol{w} \in \Omega \subseteq \mathbb{R}^d$, such that the scores of positive examples are higher than negative examples. Furthermore, when appling the score function to a dataset $\mathcal{S} \in \mathcal{X}^N$, we denote $h_{\boldsymbol{w}} : \mathcal{X}^N \mapsto \mathbb{R}^N$, where the k-th element of $h_{\boldsymbol{w}}(\mathcal{S})$ has the top-k values of $\{h_{\boldsymbol{w}}(\boldsymbol{x})|\boldsymbol{x} \in \mathcal{S}\}$. Denote the asymptotic upper bound on complexity as $\mathcal{O}$, and denote asymptotically equivalent as $\asymp$.

In this work, our main interest is to optimize a score function in the view of AUPRC:

$$\begin{aligned}\text{AUPRC}(\boldsymbol{w}; \mathcal{D}) &= \int_0^1 \mathbb{P}(y = 1|h_{\boldsymbol{w}}(\boldsymbol{x}) \geq c)\, d\,\mathbb{P}(h_{\boldsymbol{w}}(\boldsymbol{x}) \geq c|y = 1) \\ &= \int_0^1 \frac{\pi TPR(c)}{\pi TPR(c) + (1-\pi)FPR(c)}\, d\,\mathbb{P}(h_{\boldsymbol{w}}(\boldsymbol{x}) \geq c|y = 1),\end{aligned} \tag{1}$$

where $(\boldsymbol{x}, y) \sim \mathcal{D}$, c refers to a threshold, and $TPR(c) = \mathbb{P}(h_{\boldsymbol{w}}(\boldsymbol{x}) \geq c|y = 1), FPR(c) = \mathbb{P}(h_{\boldsymbol{w}}(\boldsymbol{x}) \geq c|y = 0)$. For a finite set $\mathcal{S}$, AUPRC is typically approximated by replacing the distribution function $\mathbb{P}(h_{\boldsymbol{w}}(\boldsymbol{x}) \geq c|y = 1)$ with its empirical cumulative distribution function [8, 19]:

$$\widehat{\text{AUPRC}}(\boldsymbol{w}; \mathcal{S}) = \hat{\mathbb{E}}_{\boldsymbol{x}^+ \sim \mathcal{S}^+} \left[\frac{\pi \widehat{TPR}(h_{\boldsymbol{w}}(\boldsymbol{x}^+))}{\pi \widehat{TPR}(h_{\boldsymbol{w}}(\boldsymbol{x}^+)) + (1-\pi)\widehat{FPR}(h_{\boldsymbol{w}}(\boldsymbol{x}^+))} \right], \tag{2}$$

where $\widehat{TPR}(c) = \hat{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{S}^+}\left[\ell_{0,1}(c - h_{\boldsymbol{w}}(\boldsymbol{x}))\right], \widehat{FPR}(c) = \hat{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{S}^-}\left[\ell_{0,1}(c - h_{\boldsymbol{w}}(\boldsymbol{x}))\right]$, $\ell_{0,1}(x) = 1$ if $x \leq 0$ or $\ell_{0,1}(x) = 0$ otherwise. It has been shown that $\widehat{\text{AUPRC}}$ is an unbiased estimator when $N^+/(N^+ + N^-) \to \pi$ and $N \to \infty$ [8]. With the above estimation, we have the following optimization objective:

$$\min_{\boldsymbol{w}} \quad \widehat{\text{AUPRC}}^{\downarrow}(\boldsymbol{w}; \mathcal{S}) = 1 - \widehat{\text{AUPRC}}(\boldsymbol{w}; \mathcal{S}) = \hat{\mathbb{E}}_{\boldsymbol{x}^+ \sim \mathcal{S}^+} \left[\sigma\left(\frac{1-\pi}{\pi} \cdot \frac{\widehat{FPR}(h_w(\boldsymbol{x}^+))}{\widehat{TPR}(h_w(\boldsymbol{x}^+))} \right) \right], \tag{3}$$

where $\sigma(x) = x/(1 + x)$ is concave and monotonically increasing. To make it smooth, surrogate losses ℓ_1, ℓ_2 are used to replace $\ell_{0,1}$ in $\widehat{FPR}$ and $\widehat{TPR}$ respectively, yielding the following surrogate objective:

$$\min_{\boldsymbol{w}} \quad f(\boldsymbol{w}; \mathcal{S}) = \hat{\mathbb{E}}_{\boldsymbol{x}^+ \sim \mathcal{S}^+} \left[\sigma\left(\frac{1-\pi}{\pi} \cdot \frac{\widehat{FPR}(h_w(\boldsymbol{x}^+); \ell_1)}{\widehat{TPR}(h_w(\boldsymbol{x}^+); \ell_2)} \right) \right], \tag{4}$$

where $\widehat{TPR}(c; \ell_2) = \hat{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{S}^+}\left[\ell_2(c - h_{\boldsymbol{w}}(\boldsymbol{x}))\right], \widehat{FPR}(c; \ell_1) = \hat{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{S}^-}\left[\ell_1(c - h_{\boldsymbol{w}}(\boldsymbol{x}))\right]$. Specifically, when $N^+/(N^+ + N^-) = \pi$, it is equivalent to another commonly used formulation **Average Precision (AP) Loss**:

$$\widehat{\text{AP}}^{\downarrow}(\boldsymbol{w}; \mathcal{S}) = \hat{\mathbb{E}}_{\boldsymbol{x}^+ \sim \mathcal{S}^+} \left[\sigma\left(\frac{\sum_{\boldsymbol{x} \sim \mathcal{S}^-}\left[\ell_1(h_{\boldsymbol{w}}(\boldsymbol{x}^+) - h_{\boldsymbol{w}}(\boldsymbol{x}))\right]}{\sum_{\boldsymbol{x} \sim \mathcal{S}^+}\left[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x}^+) - h_{\boldsymbol{w}}(\boldsymbol{x}))\right]} \right) \right]. \tag{5}$$

2.2 Stochastic Learning of AUPRC

Under the stochastic learning framework for instancewise losses, the empirical risk $F(\boldsymbol{w}; \mathcal{S})$ is expressed as a sum of instancewise losses: $F(\boldsymbol{w}; \mathcal{S}) = \frac{1}{N}\sum_{\boldsymbol{x} \sim \mathcal{S}} \hat{f}(\boldsymbol{w}; \boldsymbol{x})$, where $\hat{f}(\boldsymbol{w}; \boldsymbol{x})$ is the stochastic estimator of $F(\boldsymbol{w}; \mathcal{S})$. Different from instancewise losses, listwise losses like AUPRC require a batch of samples to calculate the stochastic estimator. Specifically, at each step, a subset of $\mathcal{S}$: $\boldsymbol{z} = \boldsymbol{z}^+ \cup \boldsymbol{z}^-$ is randomly drawn, where $\boldsymbol{z}^+$ consists of n^+ positive examples and $\boldsymbol{z}^-$ consists of n^- negative examples. Then a stochastic estimator of the loss function, denoted as $\hat{f}(\boldsymbol{w}; \boldsymbol{z})$, is computed with $\boldsymbol{z}$. Similar to the instancewise case, we consider a variant of the empirical/population AUPRC risks as approximations, which is a sum of stochastic losses w.r.t. all posible $\boldsymbol{z}$:

$$F(\boldsymbol{w}; \mathcal{S}) = \frac{1}{M}\sum_{\boldsymbol{z}} \hat{f}(\boldsymbol{w}; \boldsymbol{z}), \quad F(\boldsymbol{w}) = \mathbb{E}_{\mathcal{S} \sim \mathcal{D}}[F(\boldsymbol{w}; \mathcal{S})], \tag{6}$$

where M is the number of all posible $\boldsymbol{z}$. Unfortunately, due to the non-decomposability of the empirical AUPRC risk $f(\boldsymbol{w}; \mathcal{S})$, it is tackle to determine the approximation errors between $F(\boldsymbol{w}; \mathcal{S})$

and $f(\boldsymbol{w};\mathcal{S})$ in general. Nonetheless, in Sec. 3.3 we argue that by selecting proper $\hat{f}(\boldsymbol{w};\boldsymbol{z})$, $F(\boldsymbol{w};\mathcal{S})$ can be asymptotically unbiased estimator of $f(\boldsymbol{w};\mathcal{S})$, which naturally makes $F(\boldsymbol{w})$ an asymptotically unbiased estimator of $1-$ AUPRC. **In this case, $\hat{f}$ is said to be an asymptotically unbiased stochastic estimator. Moreover, if the unbiasedness holds under biased sampling rate, it is said to be sampling-rate-invariant asymptotically unbiased.**

3 Asymptotically Unbiased Stochastic AUPRC Optimization

In this section, we will present our SGD-style stochastic optimization algorithm of AUPRC. In Sec. 3.1, we propose surrogate losses to make the objective function differentiable. In Sec. 3.2, we present details of the proposed stochastic estimator and the corresponding optimization algorithm. Analyses on approximation errors are provided in Sec. 3.3.

3.1 Differentiable Surrogate Losses

Since $\ell_{0,1}$ appears in both the numerator and denominator of Eq. (4), simply implementing ℓ_1, ℓ_2 with a single function [56, 9, 54] will bring difficulty to analyze the relationship between $\widehat{\text{AUPRC}}^{\downarrow}(\boldsymbol{w};\mathcal{S})$ and $f(\boldsymbol{w};\mathcal{S})$. This motivates us to choose $\ell_1 \geq \ell_{0,1}, \ell_2 \leq \ell_{0,1}$, such that $\widehat{\text{AUPRC}}^{\downarrow}(\boldsymbol{w};\mathcal{S}) \leq f(\boldsymbol{w};\mathcal{S})$, thus the original empirical risk could be optimized by minimizing its upper bound $f(\boldsymbol{w};\mathcal{S})$. Concretely, ℓ_1 and ℓ_2 are defined as the one-side Huber loss and the one-side sigmoid loss:

$$\ell_1(x) = \begin{cases} -2x/\tau_1, & x < 0, \\ (1-x/\tau_1)^2, & 0 \leq x < \tau_1, \\ 0, & x \geq \tau_1. \end{cases} \qquad \ell_2(x) = \begin{cases} \dfrac{\exp(-x/\tau_2)-1}{\exp(-x/\tau_2)+1}, & x < 0, \\ 0, & x \geq 0. \end{cases} \tag{7}$$

Here $\tau_1, \tau_2 > 0$ are hyperparameters. ℓ_1 is convex and decreasing, which ensures the gap between positive-negative pairs is effectively optimized. Additionally, compared with the square loss and the exponential loss, ℓ_1 is more robust to noises. ℓ_2 is Lipschitz continuous, and $\ell_2 \to \ell_{0,1}$ with $\tau_2 \to 0$.

3.2 Stochastic Estimator of AUPRC

The key to a stochastic learning framework is the design of the stochastic estimator (or the corresponding gradients), *i.e.*, $\hat{f}(\boldsymbol{w};\boldsymbol{z})$. Existing methods [9, 74, 12] implement it with $\widehat{\text{AP}}^{\downarrow}(\boldsymbol{w};\boldsymbol{z})$ (Eq. (5)), which might suffer from two problems:

- **(P1)** Comparing Eq. (4) and Eq. (5), it can be seen that only when $n^+/(n^+ + n^-) \to \pi$, $\widehat{\text{AP}}^{\downarrow}$ is an asymptotically unbiased estimator. However, it is hardly satisfied since the sampling strategy is usually biased in practice.
- **(P2)** Each term in the summation of $\widehat{\text{AP}}^{\downarrow}$ is related to all instances of a batch, leading to weak leave-one-out stability, *i.e.*, changing one instance might result in a relatively large fluctuation in the stochastic gradient, especially when changing a positive example.

To tackle the above problems, we first substitute $\widehat{FPR}(h_w(\boldsymbol{x}^+);\ell_1)$ with $\hat{\mathbb{E}}_{\boldsymbol{x}\sim\boldsymbol{z}^-}[\ell_1(h_{\boldsymbol{w}}(\boldsymbol{x}^+) - h_{\boldsymbol{w}}(\boldsymbol{x}))]$, and then introduce an auxiliary vector $\boldsymbol{v} \in \mathbb{R}^{N^+}$ to estimate $\widehat{TPR}$. Formally, we propose the following batch-based estimator:

$$\hat{f}(\boldsymbol{w};\boldsymbol{z}) = \hat{f}(\boldsymbol{w};\boldsymbol{z},\boldsymbol{v}) = \hat{\mathbb{E}}_{\boldsymbol{x}^+\sim\boldsymbol{z}^+}\left[\sigma\left(\frac{1-\pi}{\pi}\cdot\frac{\hat{\mathbb{E}}_{\boldsymbol{x}\sim\boldsymbol{z}^-}[\ell_1(h_{\boldsymbol{w}}(\boldsymbol{x}^+) - h_{\boldsymbol{w}}(\boldsymbol{x}))]}{\hat{\mathbb{E}}_{v\sim\boldsymbol{v}}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x}^+) - v)]}\right)\right]. \tag{8}$$

Such an estimator enjoys two advantages: in terms of **P1**, it is asymptotically unbiased regardless of the sampling rate (see Sec. 3.3 for detailed discussions); as for **P2**, we use $\boldsymbol{v}$ to substitute $h_{\boldsymbol{w}}(\mathcal{S}^+)$, such that each positive example in a mini-batch only appears in one term. Ideally, it can be considered as using all positive examples in the dataset to estimate $\widehat{TPR}$ instead of that from a mini-batch. With the fact that $n^- \gg n^+$, this makes the corresponding algorithm more stable. Moreover, based on the model stability, generalization bounds are available (see Sec. 4).

3.3 Analyses on Approximation Errors

In this subsection, we analyze errors from two approximations in the above algorithm: **1)** the gap between $F(\boldsymbol{w};\mathcal{S})$ and the true AUPRC loss; **2)** the gap between the interpolated scores $\phi(h_{\boldsymbol{w}}(\boldsymbol{z}^+))$ and the true scores $h_{\boldsymbol{w}}(\mathcal{S}^+)$. Proofs are provided in Appendix B.1.

Denote $\pi = N^+/(N^+ + N^-)$ and $\pi_0 = n^+/(n^+ + n^-)$. We would like to show that for all $\boldsymbol{w} \in \Omega$, $\mathbb{E}_{\boldsymbol{z}}[\hat{f}(\boldsymbol{w};\boldsymbol{z})]$ is an unbiased estimator when $n \to \infty$, no matter how π_0 is chosen, while for $\mathbb{E}_{\boldsymbol{z}}[\widehat{\mathrm{AP}}^{\downarrow}(\boldsymbol{w};\boldsymbol{z})]$, it holds only when $\pi_0 = \pi$. Since only one model $\boldsymbol{w}$ is considered, we let $\boldsymbol{w}_t = \boldsymbol{w}$ in the update rule of $\boldsymbol{v}$ (Eq. (10)), and we have the following proposition:

Proposition 1. *Consider updating $\boldsymbol{v}$ with Eq. (10) for T steps, then we have*

$$\mathbb{E}[\boldsymbol{v}] = \mathbb{E}[\phi(h_{\boldsymbol{w}}(\boldsymbol{z}^+))] + (1-\beta)^T\left(\boldsymbol{v}_1 - \mathbb{E}[\phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}^+)\right)]\right), \quad Var[\boldsymbol{v}] \leq Var[\phi(h_{\boldsymbol{w}}(\boldsymbol{z}^+))] \cdot \frac{\beta}{2-\beta}.$$

Remark 1. *Two conclusions could be drawn from the above proposition: first, if the linear interpolation is asymptotically unbiased (see next subsection), by choosing a large T or setting $\boldsymbol{v}_1 = \mathbb{E}[\phi(h_{\boldsymbol{w}}(\boldsymbol{z}^+))]$, we have $\mathbb{E}[\boldsymbol{v}] \approx h_{\boldsymbol{w}}(\mathcal{S}^+)$; second,* ***by choosing a smaller β, $\boldsymbol{v}$ is more likely to concentrate on*** *$h_{\boldsymbol{w}}(\mathcal{S}^+)$.*

Proposition 2. *Assume the linear interpolation is asymptotically unbiased. Let $\kappa_1^2 = \hat{\mathbb{E}}_{c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}[Var_{\boldsymbol{x}\sim\mathcal{S}^-}[\ell_1(c - h_{\boldsymbol{w}}(\boldsymbol{x}))]]$, $\kappa_2^2 = \hat{\mathbb{E}}_{c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}[Var_{v\sim\boldsymbol{v}}[\ell_2(c-v))]]$. When $\kappa_1^2/n^- \to 0$, $\kappa_2^2/n^+ \to 0$, then there exists a positive scale H, such that*

$$\hat{\mathbb{E}}_{\boldsymbol{z}\subseteq\mathcal{S}}[\hat{f}(\boldsymbol{w};\boldsymbol{z})] \xrightarrow{P} \widehat{AUPRC}^{\downarrow}(\boldsymbol{w};\mathcal{S}), \quad \hat{\mathbb{E}}_{\boldsymbol{z}\subseteq\mathcal{S}}\left[\widehat{AP}^{\downarrow}(\boldsymbol{w};\boldsymbol{z})\right] \xrightarrow{P} (1 + (\pi_0 - \pi)H) \cdot \widehat{AUPRC}^{\downarrow}(\boldsymbol{w};\mathcal{S}),$$

where $\xrightarrow{P}$ refers to convergence in probability, and $\boldsymbol{z} \subseteq \mathcal{S}$ refers to subsets described in Sec. 2.2.

Remark 2. *The above proposition suggests that the proposed batch-based estimator is sampling-rate-invariant asymptotically unbiased, while $\widehat{AP}^{\downarrow}$ tends to be larger when the sampling rate of the positive class is greater than the prior, and vice versa. We also provide a non-asymptotic result in Appendix B.2.*

Simulation experiments are conducted as complementary to the theory. Following previous work [8], the scores are drawn from three types of distributions, including binormal, bibeta and offset uniform. The results of binormal distribution are visualized in Fig. 1, and detailed descriptions and more results are available in Appendix B.2. These results are consistent with the above remark.

Next we further study the interpolation error. For the sake of presentation, denote $p : [0,1] \mapsto \mathbb{R}$ to be an increasing score function describing $h_{\boldsymbol{w}}(\mathcal{S}^+)$, where $p(x)$ is the score in the bottom x-quantile of $h_{\boldsymbol{w}}(\mathcal{S}^+)$. Similarly, let $\hat{p}$ to be the interpolation results of $\mathbb{E}_A[h_{\boldsymbol{w}}(\boldsymbol{z}^+)]$. Assume that $\mathbb{E}_A[h_{\boldsymbol{w}}(\boldsymbol{z}^+)]$ are located in the (i/n^+)-quantiles of p, where $i \in [n^+]$, such that $p(i/n^+) = \hat{p}(i/n^+)$ and all interpolation intervals are with length $1/n^+$. The following proposition provides an upper bound of the approximation error (see [61] for proof):

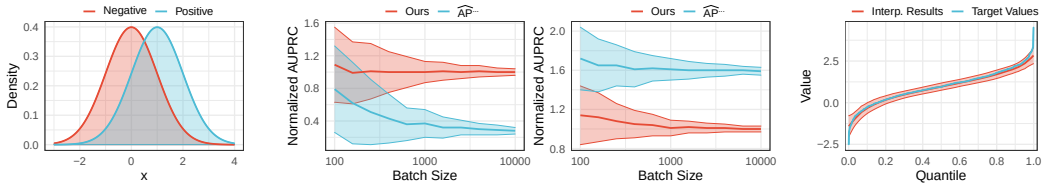
Proposition 3 (**Linear Interpolation Error**)**.** *Let $p, \hat{p}$ be defined as above. Then we have*

$$\|p - \hat{p}\|_\infty \leq \|p''\|_\infty / \left(8(n^+)^2\right).$$

Similar to the last subsection, simulation results are shown in Fig. 1(c), which shows the expected errors of linear interpolation are ignorable.

3.4 Optimization Algorithm

In the rest of this section, we focus on how to optimize $F(\boldsymbol{w};\mathcal{S})$. The main challenge is to design update rules for $\boldsymbol{v}$, such that it could efficiently and effectively approximate $h_{\boldsymbol{w}}(\mathcal{S}^+)$ without full-batch scanning. To overcome the challenge, we propose an algorithm called **Stochastic Optimization of AUPRC (SOPRC)**, which jointly updates model parameters $\boldsymbol{w}$ and the auxiliary vector $\boldsymbol{v}$. A summary of the detailed process is shown as Alg. 1. At step t, a batch of data is sampled from the training set, and then compute the corresponding scores. Afterward, scores of positive examples are mapped into a N^+-dimension vector with linear interpolation ϕ as shown in Alg. 2. $\boldsymbol{v}_{t+1}$ are updated with the interpolated scores in a moving average manner.



(a) Density functions of scores.

(b) Stochastic Estimation Errors with $\pi_0 = 0.02$ (left) and $\pi_0 = 0.2$ (right).

(c) Interpolation Errors with $\pi_0 = 0.03$.

Figure 1: Empirical analysis of estimation errors on simulation data.

Practically, n^+, n^- are finite, causing inevitable estimation errors in $f(\boldsymbol{w}; \boldsymbol{z}_{i_t}, \boldsymbol{v}_{t+1})$. Notice that another factor influencing the stochastic estimation errors, *i.e.*, κ_1^2 and κ_2^2. To reduce them, it is expected that the variance of positive (negative) scores are small, which motivates us to add a variance regularization term. However, it might force to reduce positive scores that higher than the mean value, which is contrary to our target. Therefore, we propose a **semi-variance regularization term** [4]:

$$\mathcal{L}_{var} = \frac{\lambda_1}{n^+} \sum_{\substack{\boldsymbol{x} \sim \boldsymbol{z}^+ \\ h_{\boldsymbol{w}}(\boldsymbol{x}) < \mu^+}} (h_{\boldsymbol{w}}(\boldsymbol{x}) - \mu^+)^2 + \frac{\lambda_2}{n^-} \sum_{\substack{\boldsymbol{x} \sim \boldsymbol{z}^- \\ h_{\boldsymbol{w}}(\boldsymbol{x}) > \mu^-}} (h_{\boldsymbol{w}}(\boldsymbol{x}) - \mu^-)^2, \tag{9}$$

where $\mu^+ = \frac{1}{n^+} \sum_{\boldsymbol{x} \sim \boldsymbol{z}^+} h_{\boldsymbol{w}}(\boldsymbol{x})$, $\mu^- = \frac{1}{n^-} \sum_{\boldsymbol{x} \sim \boldsymbol{z}^-} h_{\boldsymbol{w}}(\boldsymbol{x})$, λ_1, λ_2 are hyperparameters. Finally, we compute the gradients of $f(\boldsymbol{w}; \boldsymbol{z}_{i_t}, \boldsymbol{v}_{t+1}) + \mathcal{L}_{var}$, and update parameters $\boldsymbol{w}$ with gradient descent.

Algorithm 1 SOPRC

Input: Training dataset $\mathcal{S}$, maximum iterations T, learning rate $\{\eta_t\}_{t=1}^T$ and $\{\beta_t\}_{t=1}^T$.
Output: model parameters $\boldsymbol{w}_{T+1}$.

1: Initialize model parameters $\boldsymbol{w}_1$ and $\boldsymbol{v}_1$.
2: **for** $t = 1$ to T **do**
3: Sample a subset $\boldsymbol{z}_{i_t}$ from $\mathcal{S}$.
4: Compute $h_{\boldsymbol{w}_t}(\boldsymbol{z}_{i_t}^+)$ and map the results into $\phi(h_{\boldsymbol{w}_t}(\boldsymbol{z}_{i_t}^+))$ with Alg. 2.
5: Update $\boldsymbol{v}$ with

$$\begin{aligned} \boldsymbol{v}_{t+1} =& (1 - \beta_t)\boldsymbol{v}_t \\ &+ \beta_t \phi(h_{\boldsymbol{w}_t}(\boldsymbol{z}_{i_t}^+)). \end{aligned} \tag{10}$$

6: Compute $\mathcal{L}_{var}$ with Eq. (9).
7: Update the model parameter:

$$\begin{aligned} \boldsymbol{w}_{t+1} =& \boldsymbol{w}_t - \eta_t \cdot \nabla \mathcal{L}_{var} \\ &- \eta_t \cdot \nabla f(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}_{t+1}). \end{aligned} \tag{11}$$

8: **end for**

Algorithm 2 Score Interpolation $\phi(\cdot)$

Input: A real value vector $\boldsymbol{u} \in \mathbb{R}^n$ where $n < N^+$, range of target values $[b, B]$.
Output: Interpolated vector $\boldsymbol{m} = \phi(\boldsymbol{u})$.

1: Sort $\boldsymbol{u}$ in descending order.
2: Initialize $\boldsymbol{m}$ as $\mathbf{0}_{N^+}$, let $u_0 = max(2u_1 - u_2, b), u_{n+1} = min(2u_n - 2u_{n-1}, B)$.
3: **for** $i = 1$ to n **do**
4: **for** $j = \lceil \frac{N^+(i-1)}{n} \rceil$ to $\left\lceil \frac{N^+ \cdot i}{n} \right\rceil$ **do**
5:

$$\begin{aligned} m_j \mathrel{+}= \big[& (i - jn/N^+)u_{i-1} \\ & + (1 + jn/N^+ - i)u_i \big] / 2 \end{aligned}$$

6: **end for**
7: **for** $j = \left\lceil \frac{N^+ \cdot i}{n} \right\rceil$ to $\lfloor \frac{N^+ \cdot (i+1)}{n} \rfloor$ **do**
8:

$$\begin{aligned} m_j \mathrel{+}= \big[& (i + 1 - jn/N^+)u_{i-1} \\ & + (jn/N^+ - i)u_i \big] / 2 \end{aligned}$$

9: **end for**
10: **end for**

4 Generalization of SOPRC via Stability

In this section, we turn to study the *excess generalization error* of the proposed algorithm. Formally, following standard settings [5], we consider the test error of the model $A(\mathcal{S})$ trained on the training set $\mathcal{S}$. Our target is to seek an upper bound of the excess error $\mathbb{E}_{A,\mathcal{S}}[F(A(\mathcal{S})) - F(\boldsymbol{w}^*)]$, where $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w} \in \Omega} \mathbb{E}_{A,\mathcal{S}}[F(\boldsymbol{w}^*)]$. It can be decomposed as:

$$\mathbb{E}_{\mathcal{S},A}[F(A(\mathcal{S})) - F(\boldsymbol{w}^*)] = \underbrace{\mathbb{E}_{\mathcal{S},A}[F(A(\mathcal{S})) - F(A(\mathcal{S}); \mathcal{S})]}_{\textit{Estimation Error}} + \underbrace{\mathbb{E}_{\mathcal{S},A}[F(A(\mathcal{S}); \mathcal{S}) - F(\boldsymbol{w}^*)]}_{\textit{Optimization Error}}.$$

The estimation error sources from the gap of minimizing the empirical risk instead of the expected risk. In Sec. 4.1, we provide detailed discussion on the estimation error. The optimization error measures the gap between the minimum empirical risk and the results obtained by the optimization

algorithm, which will be studied in Sec. 4.2. Detailed proofs of this section are available in Appendix C. Before the formal presentation, we show the main assumptions:

Assumption 1 (Bounded Scores & Gradient). $|\hat{f}(\boldsymbol{w};\cdot)| \leq B, \|\nabla \hat{f}(\boldsymbol{w};\cdot)\|_2 \leq G$ *for all* $\boldsymbol{w} \in \Omega$.

Assumption 2 (L-Smooth Loss). $\|\nabla \hat{f}(\boldsymbol{w};\cdot) - \nabla \hat{f}(\tilde{\boldsymbol{w}};\cdot)\|_2 \leq L\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|_2$ *for all* $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \Omega$.

Assumption 3 (Lipschitz Continuous Functions). $|\ell_1(x) - \ell_1(\tilde{x})| \leq L_1|x - \tilde{x}|$, $|\ell_2(x) - \ell_2(\tilde{x})| \leq L_2|x - \tilde{x}|$ *for all* $x, \tilde{x} \in [-2B, 2B]$. $\|\phi(\boldsymbol{x}) - \phi(\tilde{\boldsymbol{x}})\|_2 \leq C_\phi\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2$ *for all* $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^{N^+}$.

4.1 Generalization of AUPRC via Model Stability

The generalization of SGD-style algorithms for instancewise loss has been widely studied with stability measure [39, 21, 28]. However, these results could not be directly applied to listwise losses like AUPRC. The main reason is that the estimation of each stochastic gradient requires a list of examples, and the estimation is usually biased. Nonetheless, to bridge the optimization algorithm and the generalization of AUPRC, we propose a listwise variant of *on-average model stability* [39] as follows:

Definition 1 (Listwise On-average Model Stability). *Let* $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ *and* $\widetilde{\mathcal{S}} = \{(\widetilde{\boldsymbol{x}}_i, y_i)\}_{i=1}^N$ *be two sets of examples whose features are drawn independently from* $\mathcal{X}$. *For any* $i = 1, \cdots, N$, *denote* $\mathcal{S}^{(i)} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_{i-1}, y_{i-1}), (\widetilde{\boldsymbol{x}}_i, y_i), (\boldsymbol{x}_{i+1}, y_{i+1}), \cdots, (\boldsymbol{x}_n, y_n)\}$. *A stochastic algorithm* A *is listwise on-average model* (ϵ^+, ϵ^-)*-stable if the following condition holds:*

$$\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},A}\left[\frac{1}{N^+}\sum_{y_i=1}\left\| A(\mathcal{S}) - A(\mathcal{S}^{(i)})\right\|_2\right] \leq \epsilon^+, \mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},A}\left[\frac{1}{N^-}\sum_{y_i=-1}\left\| A(\mathcal{S}) - A(\mathcal{S}^{(i)})\right\|_2\right] \leq \epsilon^-.$$

The following theorem shows that the estimation error is bounded by the above-defined stability:

Theorem 1 (Generalization via Model Stability). *Let a stochastic algorithm* A *be listwise on-average model* (ϵ^+, ϵ^-)*-stable and Asmp. 1 holds. Then we have*

$$\mathbb{E}_{\mathcal{S},A}\left[F(A(\mathcal{S})) - F(A(\mathcal{S});\mathcal{S})\right] \leq G(n^+\epsilon^+ + n^-\epsilon^-). \tag{12}$$

With the above theorem, now we only need to focus on the model stability of the proposed algorithm. Notice that in Alg. 1, both $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are updated at each step, thus we have to consider the stability of both simultaneously. The following lemma provides a recurrence for the stability $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$.

Lemma 1. *Let* $\mathcal{S}, \widetilde{\mathcal{S}}, \mathcal{S}^{(i)}$ *be constructed as Def. 1 and Asmp. 1, 2, 3 hold. Let* $\{\boldsymbol{w}_t\}_t$ *and* $\{\boldsymbol{w}_t^{(i)}\}_t$ *be produced by Alg. 1 with* $\mathcal{S}$ *and* $\mathcal{S}^{(i)}$, *respectively. Denote* $L = \max\{L_w, L_v/n^+, C_\phi B, G/2, B'_\ell\}$, $\boldsymbol{m}_t^{(i)} = \begin{bmatrix} \|\boldsymbol{w}_t - \boldsymbol{w}_t^{(i)}\|_2 & \|\boldsymbol{v}_t - \boldsymbol{v}_t^{(i)}\|_2 & 1 \end{bmatrix}^\top$, $\boldsymbol{m}_t^+ = \frac{1}{N^+}\sum_{y_i=1}\mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}_t^{(i)}\right]$, $\boldsymbol{m}_t^- = \frac{1}{N^+}\sum_{y_i=-1}\mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}_t^{(i)}\right]$. *Then for all* $t \in [T]$, *by setting* $\beta_t \leq 2C_\phi B/n^+$, *we have*

$$\boldsymbol{m}_{t+1}^+ \leq \left(\boldsymbol{I}_3 + \boldsymbol{R}_t^+\right)\cdot \boldsymbol{m}_t^+, \qquad \boldsymbol{m}_{t+1}^- \leq \left(\boldsymbol{I}_3 + \boldsymbol{R}_t^-\right)\cdot \boldsymbol{m}_t^-, \tag{13}$$

where $\boldsymbol{I}_3$ *is the* 3×3 *identity matrix and*

$$R_t^+ = \begin{bmatrix} 2L\eta_t & \frac{L(1-\beta_t)\eta_t}{N^+} & \frac{L\eta_t}{N^+} \\ L\beta_t & 0 & \frac{1}{N^+} \\ 0 & 0 & 0 \end{bmatrix}, R_t^- = \begin{bmatrix} 2L\eta_t & \frac{L_v(1-\beta_t)\eta_t}{N^+} & \frac{L\eta_t\cdot n^+}{N^-} \\ L\beta_t & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{14}$$

Finally, we utilize the matrix spectrum of R_t^+ and R_t^- to show that the model stability w.r.t. Alg. 1 decreases as the number of training examples increases (see Appendix C.2 for details):

Theorem 2. *Let* $\lambda = LC_\eta(1 + \sqrt{1 - \beta^2 + \beta})$, *and assumptions in Lem. 1 hold. By setting* $\eta_t \leq \frac{C_\eta}{t}$, $\beta_t = \beta \asymp 1/n^+$ *and* $T \leq N^+$, *Alg. 1 is list on-average model stable with*

$$\epsilon^+ = \mathcal{O}\left(\frac{(Tn^+)^{\frac{\lambda}{\lambda+1}}}{N^+}\right), \epsilon^- = \mathcal{O}\left(\frac{(Tn^-)^{\frac{\lambda}{\lambda+1}}}{N^-}\right). \tag{15}$$

Table 1: Quantitative results on SOP, iNaturalist, and VehicleID. All methods are trained with training sets. The best and the second best results are highlighted in soft red and soft blue, respectively.

Methods	Stanford Online Products			iNaturalist			PKU VehicleID		
	mAUPRC	R@1	R@10	mAUPRC	R@1	R@4	mAUPRC	R@1	R@5
Contrastive loss [27]	57.73	77.60	89.31	27.99	54.19	71.12	67.26	87.46	94.60
Triplet loss [32]	58.07	78.34	90.50	30.59	60.53	77.62	70.99	90.09	95.54
MS loss [71]	60.10	79.64	90.38	30.28	63.39	78.50	69.15	88.82	95.06
XBM [72]	61.29	80.66	91.08	27.46	59.12	75.18	71.24	**92.78**	95.83
SmoothAP [9]	**61.65**	**81.13**	**92.02**	**33.92**	**66.13**	**80.93**	72.28	91.31	96.05
DIR [58]	60.74	80.52	91.35	33.51	64.86	79.79	**72.72**	91.38	**96.10**
FastAP [12]	57.10	77.30	89.61	31.02	56.64	73.57	70.82	89.42	95.38
AUROC [25]	55.80	77.32	89.64	27.24	60.88	77.76	58.12	81.73	91.92
BlackBox [52]	59.74	79.48	90.74	29.28	56.88	74.10	70.92	90.14	95.52
Ours	**62.75**	**81.91**	**92.50**	**36.16**	**68.22**	**82.86**	**74.92**	**92.56**	**96.43**

4.2 Convergence of AUPRC Stochastic Optimization

Following previous work [24, 34], we study the optimization error of the proposed algorithm under the *Polyak-Łojasiewicz (PL)* condition. It has been shown that the PL condition holds for several widely used models including some classes of neural networks [13, 42].

Assumption 4 (Polyak-Łojasiewicz Condition [34, 37]). *Denote $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\Omega} F(\boldsymbol{w})$. Assume F satisfy the expectation version of PL condition with parameter $\mu > 0$, i.e.,*

$$\mathbb{E}_{\mathcal{S}}[F(\boldsymbol{w};\mathcal{S}) - F(\boldsymbol{w}^*)] \leq \frac{1}{\mu}\mathbb{E}_{\mathcal{S}}[\|\nabla F(\boldsymbol{w};\mathcal{S})\|_2^2]. \tag{16}$$

The main difference to the existing convergence analysis on non-convex optimization is that the gradient estimation is biased. Nonetheless, we show that the bias terms from Alg. 1 tend to 0 with sufficient training data and training time (see Appendix C.3), leading to the following convergence:

Theorem 3. *Let Asmp. 1, 3, 4 hold. By setting $\eta_t = \frac{2t+1}{\mu(t+1)^2}$ and $\beta_t = \beta \asymp 1/n^+$, we have*

$$\mathbb{E}_A[F(\boldsymbol{w}_{T+1}) - F(\boldsymbol{w}^*)] = \mathcal{O}\left(n^+/T + 1/N^+\right). \tag{17}$$

Theorem 4. *Let assumptions in Thm. 2 and 3 hold. By setting $T \asymp (N^+)^{\frac{\lambda+1}{2\lambda+1}}(n^+)^{-\frac{1}{2\lambda+1}}$, we have*

$$\mathbb{E}_{\mathcal{S},A}[F(A(\mathcal{S})) - F(\boldsymbol{w}^*)] = \mathcal{O}\left((N^+)^{-\frac{\lambda+1}{2\lambda+1}} \cdot (n^+)^{\frac{3\lambda+1}{2\lambda+1}}\right) + \mathcal{O}\left((N^-)^{-\frac{\lambda+1}{2\lambda+1}} \cdot (n^-)^{\frac{3\lambda+1}{2\lambda+1}}\right). \tag{18}$$

Remark 3. *Recall that $\lambda = LC_\eta(1 + \sqrt{1-\beta^2+\beta})$ and $C_\eta = 4/\mu$, when β is small, we have $\lambda \approx 4L/\mu$. Here L/μ is a condition number determined by the model and surrogate losses. Notice that $n^+ \ll N^+, n^- \ll N^-$, if $\lambda = 1$, the generalization bound is $\mathcal{O}\left((N^+)^{-2/3}\cdot(n^+)^{4/3} + (N^-)^{-2/3}\cdot(n^-)^{4/3}\right)$. As λ increases, it increases to $\mathcal{O}\left((N^+)^{-1/2}\cdot(n^+)^{3/2} + (N^-)^{-1/2}\cdot(n^-)^{3/2}\right)$.*

5 Experiments

To validate the effectiveness of the proposed method, we conduct empirical studies on the image retrieval task, in which data distributions are largely skewed and AUPRC is commonly used as an evaluation metric. More detailed experimental settings are provided in Appendix D.1. The source code is available in `https://github.com/KID-7391/SOPRC.git`.

5.1 Implementation Details

Datasets. We evaluate the proposed method on three image retrieval benchmarks with various domains and scales, including **Stanford Online Products (SOP)**[2] [48], **PKU VehicleID**[3] [43] and

[2] `https://github.com/rksltnl/Deep-Metric-Learning-CVPR16`. Licensed MIT.

[3] `https://www.pkuml.org/resources/pku-vehicleid.html`. Data files © Original Authors.

iNaturalist[4] [68]. We follow the official setting to split a test set from each dataset, and then further split the rest into a training set and a validation set by a ratio of $9:1$.

Network Architecture. The feature extractor is implemented with ResNet-50 [29] pretrained on ImageNet [60]. Following previous work [12, 9], the batch normalization layers are fixed during training, and the output embeddings are mapped to 512-d with a linear projection. Given L_2 normalized embeddings of a query image $\boldsymbol{e}_q$ and a gallery list $\{\boldsymbol{e}_i\}_i$, the scores are represented by the cosine similarity $\boldsymbol{e}_q^\top \boldsymbol{e}_i$ for all i.

Optimization Strategy. In the training phase, the input images are resized such that the sizes of the shorter sides are 256. Afterward, we applied standard data augmentations including random cropping (224×224) and random flipping (50%). The model parameters are optimized in an end-to-end manner as shown in Alg. 1, where $\beta = 0.001$ and the weight decay is set to 4×10^{-4}. The default batch size is set to 224, where each mini-batch is randomly sampled such that there are exactly 4 positive examples per category. The learning rates are tuned according to performance on validation sets: for SOP, the learning rate are initialized as 0.01 and decays by 0.1 at the $15k$ and $30k$ iterations, $T = 50k$; for VehicleID, the learning rate are initialized as 0.001 and decays by 0.1 at the $40k$ and $80k$ iterations, $T = 100k$; for iNaturalist, the learning rate are initialized as 0.001 and decays by 0.1 at the $80k$ and $110k$ iterations, $T = 130k$.

Competitors. We compare two types of competitors: **1) Pairwise Losses**, including *Contrastive Loss* [27], *Triplet Loss* [32], *Multi-Similarity (MS) Loss* [71], *Cross-Batch Memory (XBM)* [72]. These methods construct loss functions with image pairs or triplets. **2) Ranking-Based Losses**, including *SmoothAP* [9], *FastAP* [12], *DIR* [58], *BlackBox* [33], and *Area Under the ROC Curve Loss (AUROC)* [76]. These methods directly optimize the ranking-based metrics.

Evaluation Metrics. In all experiments, we adopt evaluation metrics: *mean AUPRC (mAUPRC)* and *Recall@k*. mAUPRC is also called mean average precision (mAP) in literature, which takes the mean value over the AUPRC of all queries. Recall@k measures the probability that at least one positive example is ranked in the top-k list.

5.2 Main Results

We evaluate all methods with *mean AUPRC (mAUPRC)* and *Recall@k*. mAUPRC measures the mean value of the AUPRC over all queries, a.k.a. mean average precision (mAP). The performance comparisons on test sets are shown in Tab. 1. Consequently, we have the following observations: **1)** In all datasets, the proposed method surpasses all competitors in the view of mAUPRC, especially in the large-scale long-tailed dataset iNaturalist. This validates the advantages of our method in boosting the AUPRC of models. **2)** Compared to pairwise losses, the AUPRC/AP optimization methods enjoy better performance generally. The main reason is that pairwise losses could only optimize models indirectly by constraining relative scores between positive and negative example pairs, while ignoring the overall ranking. **3)** Although some pairwise methods like XBM have a satisfying performance on Recall@1, their mAUPRC is relatively low. It is caused by the limitation of Recall@1, *i.e.*, it focuses on the top-1 score while ignoring the ranking of other examples. What's more, this phenomenon shows the inconsistency of Recall@k and AUPRC, revealing the necessity of studying AUPRC optimization. More results are available in Appendix D.2.

To qualitatively demonstrate the effect of the proposed method, we also show the mean PR curves and convergence curves in Fig. 2. The left two subfigures demonstrate that the proposed method achieves can effectively improve AUPRC. The right subfigure shows the affect of batch size, from which can be seen that a large batch size leads to better performance. One of the reasons is that a small batch size will amplify the AUPRC stochastic estimation error. Such a problem have been addressed by maintaining inner gradient estimations in a moving average manner [54, 70, 75]. Unfortunately, when applied to image retrieval problems, it needs to maintain intermediate variables for each sample pair, which will bring high complexity in time and space, thus we leave this problem as further work.

5.3 Ablation Studies

We further investigate the effect of different components of the proposed method. Results are shown in Tab. 2, and more detailed statements and analyses are as follows.

[4] https://github.com/visipedia/inatcomp/tree/master/2018. Licensed MIT.

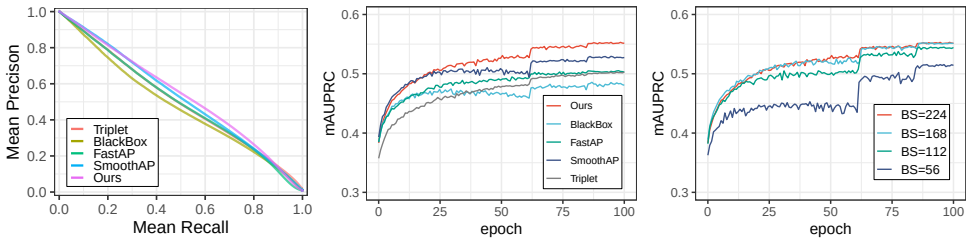


Figure 2: Qualitative results on iNaturalist. Left most: mean PR curves of different methods. Right two: convergence of different methods and batch sizes in terms of mAUPRC in the validation set.

Table 2: Ablation study over different components of our method on iNaturalist.

No.	Unb. Est.	with $\boldsymbol{v}_t$	with $\mathcal{L}_{var}$	Opt.	mAUPRC	R@1	R@4	R@16	R@32
1	✗	✗	✗	SGD	34.58	66.35	81.04	89.80	92.72
2	✓	✗	✗	SGD	35.84	67.08	81.68	90.17	92.98
3	✓	✓	✗	SGD	35.99	67.50	82.03	90.44	93.26
4	✓	✓	✓	SGD	**36.16**	**68.22**	**82.86**	**91.02**	**93.71**
5	✓	✓	✓	Adam	**36.20**	**68.48**	**82.70**	**90.96**	**93.63**

Effect of Unbiased Estimator. To show the performance drop caused by the biased estimator, we replace the prior π in Eq. (8) with $n^+/(n^+ + n^-)$. Comparing line 1 and line 2, using the unbiased estimator increases the mAUPRC by 1.3%, which is consistent with our theoretical results in Sec. 3.3. Notably, the unbiased estimator is the main source of improvements in terms of mAUPRC.
Effect of $\boldsymbol{v}_t$. To show the effect of introducing $\boldsymbol{v}_t$ to estimate $\phi(\mathcal{S}^+)$, we directly use $\phi(\boldsymbol{z}^+)$ instead in the first two lines. Comparing line 2 and line 3, using $\boldsymbol{v}_t$ could bring consistent improvements due to the better generalization ability.
Effect of $\mathcal{L}_{var}$. We show that shrinking variances could reduce the batch-based estimation errors. Comparing line 3 and line 4, it can be seen that $\mathcal{L}_{var}$ further boosts the proposed method.
Effect of Optimizer. Comparing line 4 and line 5, it can be seen that the choice of optimizer only has a slight influence.

6 Conclusion & Future Work

In this paper, we present a stochastic learning framework for AUPRC optimization. To begin with, we propose a stochastic AUPRC optimization algorithm based on an asymptotically unbiased stochastic estimator. By introducing an auxiliary vector to approximate the scores of positive examples, the proposed algorithm is more stable. On top of this, we study algorithm-dependent generalization. First, we propose list model stability to handle listwise losses like AUPRC, and bridge the generalization and the stability. Afterward, we show that the proposed algorithm is stable, leading to an upper bound of the generalization error. Experiments on three benchmarks validate the advantages of the proposed framework. One limitation is the convergence rate is controlled by the scale of the dataset. In the further, we will consider techniques like variance reduction to improve the convergence rate, and jointly consider the corresponding algorithm-dependent generalization.

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, 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, in part by the China National Postdoctoral Program for Innovative Talents under Grant BX2021298, and in part by China Postdoctoral Science Foundation under Grant 2022M713101.

References

[1] Shilong Bao, Qianqian Xu, Ke Ma, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Collaborative preference embedding against sparse labels. In *ACM International Conference on Multimedia*, pages 2079–2087, 2019.

[2] Shilong Bao, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Rethinking collaborative metric learning: Toward an efficient alternative without negative sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[3] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[4] Shaun A Bond and Stephen E Satchell. Statistical properties of the sample semi-variance. *Applied Mathematical Finance*, 9(4):219–239, 2002.

[5] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. *Advances in Neural Information Processing Systems*, 20, 2007.

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[7] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[8] Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: point estimates and confidence intervals. In *ECML PKDD*, pages 451–466. Springer, 2013.

[9] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *European Conference on Computer Vision*, pages 677–694. Springer, 2020.

[10] Christopher Burges, Robert Ragno, and Quoc Le. Learning to rank with nonsmooth cost functions. *Advances in Neural Information Processing Systems*, 19:193–200, 2006.

[11] Christopher JC Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81, 2010.

[12] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1861–1870, 2019.

[13] Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754. PMLR, 2018.

[14] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. Towards accurate one-stage object detection with ap-loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5119–5127, 2019.

[15] Kean Chen, Weiyao Lin, John See, Ji Wang, Junni Zou, et al. Ap-loss for accurate one-stage object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[16] Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. On sampling strategies for neural network-based collaborative filtering. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–776, 2017.

[17] Wei Chen, Tie-Yan Liu, Yanyan Lan, Zhi-Ming Ma, and Hang Li. Ranking measures and loss functions in learning to rank. *Advances in Neural Information Processing Systems*, 22:315–323, 2009.

[18] Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

[19] Stéphan Clémençon and Nicolas Vayatis. Nonparametric estimation of the precision-recall curve. In *International Conference on Machine Learning*, pages 185–192, 2009.

[20] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *International Conference on Machine Learning*, pages 233–240, 2006.

[21] Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

[22] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Sodeep: a sorting deep net to learn ranking loss surrogates. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10792–10801, 2019.

[23] Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[25] Wei Gao and Zhi-Hua Zhou. On the consistency of auc pairwise optimization. In *International Conference on Machine Learning*, 2015.

[26] Mark Goadrich, Louis Oliphant, and Jude Shavlik. Gleaner: Creating ensembles of first-order clauses to improve recall-precision curves. *Machine Learning*, 64(1-3):231–261, 2006.

[27] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006.

[28] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[30] Kun He, Fatih Cakir, Sarah Adel Bargal, and Stan Sclaroff. Hashing as tie-aware learning to rank. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4023–4032, 2018.

[31] Paul Henderson and Vittorio Ferrari. End-to-end training of object class detectors for mean average precision. In *Asian Conference on Computer Vision*, pages 198–213. Springer, 2016.

[32] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015.

[33] Qijia Jiang, Olaoluwa Adigun, Harikrishna Narasimhan, Mahdi Milani Fard, and Maya Gupta. Optimizing black-box metrics with adaptive surrogates. In *International Conference on Machine Learning*, pages 4784–4793. PMLR, 2020.

[34] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[35] Joon-myoung Kwon, Youngnam Lee, Yeha Lee, Seungwoo Lee, and Jinsik Park. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *Journal of the American Heart Association*, 7(13):e008678, 2018.

[36] Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. *Advances in Neural Information Processing Systems*, 33:21236–21246, 2020.

[37] Yunwen Lei, Mingrui Liu, and Yiming Ying. Generalization guarantee of sgd for pairwise learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[38] Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186. PMLR, 2021.

[39] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819, 2020.

[40] Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2019.

[41] Zhuo Li, Weiqing Min, Jiajun Song, Yaohui Zhu, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Rethinking the optimization of average precision: Only penalizing negative instances before positive ones is enough. *arXiv preprint arXiv:2102.04640*, 2021.

[42] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

[43] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016.

[44] Donald Metzler and W Bruce Croft. A markov random field model for term dependencies. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 472–479, 2005.

[45] Pritish Mohapatra, CV Jawahar, and M Pawan Kumar. Efficient optimization for average precision svm. *Advances in Neural Information Processing Systems*, 27:2312–2320, 2014.

[46] Pritish Mohapatra, Michal Rolinek, CV Jawahar, Vladimir Kolmogorov, and M Pawan Kumar. Efficient optimization for rank-based loss functions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3701, 2018.

[47] Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638. PMLR, 2018.

[48] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016.

[49] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. A ranking-based, balanced loss function unifying classification and localisation in object detection. In *Advances in Neural Information Processing Systems*, 2020.

[50] Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, 68(8):855–859, 2015.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[52] Marin Vlastelica Pogančić, Anselm Paulus, Vit Musil, Georg Martius, and Michal Rolinek. Differentiation of blackbox combinatorial solvers. In *International Conference on Learning Representations*, 2019.

[53] Tomaso Poggio and Christian R Shelton. On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49, 2002.

[54] Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in Neural Information Processing Systems*, 34, 2021.

[55] Tao Qin, Tie-Yan Liu, and Hang Li. A general approximation framework for direct optimization of information retrieval measures. *Information Retrieval*, 13(4):375–397, 2010.

[56] Tao Qin, Xu-Dong Zhang, Ming-Feng Tsai, De-Sheng Wang, Tie-Yan Liu, and Hang Li. Query-level loss functions for information retrieval. *Information Processing & Management*, 44(2):838–855, 2008.

[57] Vijay Raghavan, Peter Bollmann, and Gwang S Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, 7(3):205–229, 1989.

[58] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *International Conference on Computer Vision*, pages 5107–5116, 2019.

[59] William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[61] Timothy Sauer. *Numerical analysis*. Addison-Wesley Publishing Company, 2011.

[62] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

[63] Yang Song, Alexander Schwing, Raquel Urtasun, et al. Training deep neural networks via direct loss minimization. In *International Conference on Machine Learning*, pages 2169–2177. PMLR, 2016.

[64] Viet-Anh Tran, Romain Hennequin, Jimena Royo-Letelier, and Manuel Moussallam. Improving collaborative metric learning with efficient negative sampling. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1201–1204, 2019.

[65] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[66] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[67] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[68] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[69] Bokun Wang and Tianbao Yang. Finite-sum coupled compositional stochastic optimization: Theory and applications. In *International Conference on Machine Learning*, pages 23292–23317. PMLR, 2022.

[70] Guanghui Wang, Ming Yang, Lijun Zhang, and Tianbao Yang. Momentum accelerates the convergence of stochastic auprc maximization. In *International Conference on Artificial Intelligence and Statistics*, pages 3753–3771. PMLR, 2022.

[71] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

[72] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2020.

[73] Zitai Wang, Qianqian Xu, Ke Ma, Yangbangyan Jiang, Xiaochun Cao, and Qingming Huang. Adversarial preference learning with pairwise comparisons. In *ACM International Conference on Multimedia*, pages 656–664, 2019.

[74] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *International Conference on Machine Learning*, pages 1192–1199, 2008.

[75] Tianbao Yang. Algorithmic foundation of deep x-risk optimization. *arXiv preprint arXiv:2206.00439*, 2022.

[76] Zhiyong Yang, Qianqian Xu, Shilong Bao, Xiaochun Cao, and Qingming Huang. Learning with multiclass auc: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[77] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. A support vector method for optimizing average precision. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271–278, 2007.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see Sec. 6.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We haven't found any negative societal impact of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Because it would be too computationally expensive.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] Assets we used are open source.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

A Related Work **17**

A.1 AUPRC Optimization 17

A.2 Generalization via Stability of Stochastic Optimization 17

B Details on Error Analysis **18**

B.1 Proofs on the Error Bounds 18

B.2 Details of Simulation Experiments 21

C Proofs on Generalization via Stability **23**

C.1 Generalization by List Model Stability 23

C.2 List Model Stability of SOPRC 24

C.3 Convergence Rate 29

D Experiments **33**

D.1 Implementation Details 33

D.2 More Results 34

A Related Work

A.1 AUPRC Optimization

To measure model performances in largely skewed datasets, Raghavan *et al.*[57] use Precision-Recall (PR) curves to describe the trade-off between precision and recall, leading to a metric named Area Under the PR Curve (AUPRC). Benefiting from its insensitivity to data distribution, AUPRC has been widely used in imbalanced scenarios, such as information retrieval [17, 56, 44], recommendation systems [64, 16] and computer visions [9, 15, 12]. The important application value of AUPRC and the inconsistency with other metrics [20] have raised a wave of research on direct optimization of AUPRC. Early work can be roughly divided into discrete methods and continuous methods. The first technical route utilizes discrete optimization methods to directly optimize the non-differentiable objective, *e.g.*, Markov random field model [44], randomized search [26], dynamic programming [77], and error driven method [10]. The second technical route seeks for continuous surrogate objective, like convex upper bound on AUPRC for support vector machines (SVMs) [45]. Unfortunately, limited by the high computational complexity, these methods are not suitable for deep learning.

With the increasing application of deep learning in ranking problems, stochastic optimization for AUPRC has attracted the attention of researchers. It is challenging since AUPRC is neither differentiable nor decomposable. Therefore, the mainstream approaches tackle stochastic optimization of AUPRC from two aspects: surrogate loss functions and batch-based estimators. In the first aspect, some methods replace the non-differentiable 0-1 loss with surrogate functions like exponential loss [56], sigmoid loss [9], and linear interpolation function [33]. Contrary to approximating the objective with differentiable functions, another route tackle this problem in an error driven style. More specifically, Burges *et al.*[10, 11] propose to decompose the gradients with chain rule, and then replace the differential w.r.t. prediction scores with the difference. Since the differential of scores w.r.t. model parameters are available, the gradients can be obtained in this way. This idea has been extended to solve the imbalance problem in object detection [15, 14, 49]. However, these methods cannot guarantee the relationship between the surrogate optimization objective and the original AUPRC. In contrast, we propose a differentiable upper bound of AUPRC as a surrogate objective function.

As for the batch-based estimators, the mainstream approaches use a batch of examples to calculate average precision [9, 54, 55], or approximate precision and recall [58, 12, 30] with the histogram binning technique [65]. However, due to the biased sampling, these estimators are not asymptotically unbiased, leading to biased stochastic gradient estimations. Moreover, since the number of positive examples sampled in a batch is typically limited, these algorithms are less stable, which is not conducive to the generalization.

Limitations of existing work motivate us to design a more appropriate estimator that (asymptotically) unbiased and stable. In this work, we propose an unbiased estimator and further enhance the stability with an auxiliary set assisting the TPR estimation.

A.2 Generalization via Stability of Stochastic Optimization

The algorithmic stability [59, 7, 21] is a standard framework for generalization analysis. Besides the original uniform stability [7], various types of stability are studied, *e.g.*, expected stability [62], hypothesis set stability [23], and on-average model stability [39]. Unlike another standard framework built on the Rademacher complexity [3, 53, 66], generalization via stability takes the optimization algorithm into account. This is equivalent to constraining the hypothesis set to the possible optimization outcomes, which are usually in the neighborhood of the (local) optimum. On the other hand, this framework can naturally consider two factors simultaneously, *i.e.*, optimization errors and estimation errors, allowing jointly consideration of generalization and convergence trade-offs [18]. These advantages enable stability to be applied in a wide variety of conditions [40, 47, 13].

However, existing techniques focus on instancewise or pairwise [37, 36] loss functions, while AUPRC is a listwise metric. In addition, each term in the AUPRC are related to all instances, thus the stability is limited by the capacity of a batch. Nonetheless, we propose a listwise variant of on-average model stability [39], and further develop generalization guarantees of our proposed stochastic optimization algorithm for AUPRC.

B Details on Error Analysis

B.1 Proofs on the Error Bounds

Proposition 1. *Consider updating $\boldsymbol{v}$ with Eq. (10) for T steps, then we have*

$$\mathbb{E}[\boldsymbol{v}] = \mathbb{E}[\phi(h_{\boldsymbol{w}}(\boldsymbol{z}^+))] + (1-\beta)^T \left(\boldsymbol{v}_1 - \mathbb{E}[\phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}^+)\right)]\right), \;\; Var[\boldsymbol{v}] \leq Var[\phi(h_{\boldsymbol{w}}(\boldsymbol{z}^+))] \cdot \frac{\beta}{2-\beta}.$$

Proof. With the update rule we have

$$\boldsymbol{v} = \sum_{t=1}^{T} \beta(1-\beta)^{t-1} \phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}_{i_t}^+)\right) + (1-\beta)^T \boldsymbol{v}_1, \tag{19}$$

thus we have the expectation of $\boldsymbol{v}$

$$\begin{aligned}
\mathbb{E}[\boldsymbol{v}] =& \sum_{t=1}^{T} \beta(1-\beta)^{t-1} \mathbb{E}[\phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}^+)\right)] + (1-\beta)^T \boldsymbol{v}_1 \\
=& \beta \cdot \frac{1-(1-\beta)^T}{1-(1-\beta)} \mathbb{E}[\phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}^+)\right)] + (1-\beta)^T \boldsymbol{v}_1 \\
=& \mathbb{E}[\phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}^+)\right)] + (1-\beta)^T \left(\boldsymbol{v}_1 - \mathbb{E}[\phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}^+)\right)]\right),
\end{aligned} \tag{20}$$

and the variance

$$\begin{aligned}
Var[\boldsymbol{v}] =& \sum_{t=1}^{T} \beta^2 (1-\beta)^{2t-2} Var[\phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}^+)\right)] \\
\leq& Var[\phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}^+)\right)] \cdot \frac{\beta^2}{1-(1-\beta)^2} \\
=& Var[\phi\left(h_{\boldsymbol{w}}(\boldsymbol{z}^+)\right)] \cdot \frac{\beta}{2-\beta}
\end{aligned} \tag{21}$$

□

Proposition 2. *Assume the linear interpolation is asymptotically unbiased. Let $\kappa_1^2 = \hat{\mathbb{E}}_{c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}[Var_{\boldsymbol{x}\sim\mathcal{S}^-}[\ell_1(c - h_{\boldsymbol{w}}(\boldsymbol{x}))]]$, $\kappa_2^2 = \hat{\mathbb{E}}_{c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}[Var_{v\sim\boldsymbol{v}}[\ell_2(c - v))]]$. When $\kappa_1^2/n^- \to 0$, $\kappa_2^2/n^+ \to 0$, then there exists a positive scale H, such that*

$$\hat{\mathbb{E}}_{\boldsymbol{z}\subseteq\mathcal{S}}[\hat{f}(\boldsymbol{w};\boldsymbol{z})] \xrightarrow{P} \widehat{AUPRC}^{\downarrow}(\boldsymbol{w};\mathcal{S}), \;\; \hat{\mathbb{E}}_{\boldsymbol{z}\subseteq\mathcal{S}}\left[\widehat{AP}^{\downarrow}(\boldsymbol{w};\boldsymbol{z})\right] \xrightarrow{P} (1 + (\pi_0 - \pi)H) \cdot \widehat{AUPRC}^{\downarrow}(\boldsymbol{w};\mathcal{S}),$$

where $\xrightarrow{P}$ refers to convergence in probability, and $\boldsymbol{z} \subseteq \mathcal{S}$ refers to subsets described in Sec. 2.2.

Proof. With sufficient large N^+ and N^-, we consider $\mathcal{S}$ as the population. Given a threshold $c \in \mathbb{R}$, we consider $\ell_1(c - h_{\boldsymbol{w}}(\boldsymbol{x}))$ as i.i.d. variables controlled by $\boldsymbol{x}$ with mean $\mu_{c,1}$ and variance $\kappa_{c,1}^2$. In this way, the surrogate $\widehat{FPR}$ can be viewed as an average of these variables:

$$X_{n^-}^c = \hat{\mathbb{E}}_{\boldsymbol{x}\sim\boldsymbol{z}^-}\left[\ell_1(c - h_{\boldsymbol{w}}(\boldsymbol{x}))\right].$$

According to the central limit theorem, when $n^- \to \infty$ we have

$$X_{n^-}^c \rightsquigarrow \mathcal{N}(\mu_{c,1}, \kappa_{c,1}^2/n^-), \tag{22}$$

where $\rightsquigarrow$ refers to convergence in law. Similarly, consider $\ell_2(c - v))$ as variables with mean $\mu_{c,2}$ and standard deviation $\kappa_{c,2}^2$, and denote

$$Y_{n^+}^c = \hat{\mathbb{E}}_{v\sim\boldsymbol{v}}\left[\ell_2(c - v)\right]. \tag{23}$$

According to the portmanteau lemma [67] and Prop. 3, when $n^+ \to \infty$, we have

$$Y_{n^+}^c \rightsquigarrow \mathcal{N}(\mu_{c,2}, \kappa_{c,2}^2/n^+). \tag{24}$$

Decompose the difference between $\hat{f}(\boldsymbol{w};\boldsymbol{z})$ and $\widehat{\text{AUPRC}}^{\downarrow}(\boldsymbol{w};\mathcal{S})$ as

$$
\begin{aligned}
&\hat{\mathbb{E}}_{\boldsymbol{z}\subseteq\mathcal{S}}[\hat{f}(\boldsymbol{w};\boldsymbol{z})] - \widehat{\text{AUPRC}}^{\downarrow}(\boldsymbol{w};\mathcal{S}) \\
=& \hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\frac{(1-\pi)X^c_{n^-}}{(1-\pi)X^c_{n^-}+\pi Y^c_{n^+}}\right] - \mathbb{E}_{c\sim h_{\boldsymbol{w}}(\mathcal{S}^+)}\left[\frac{(1-\pi)\mu_{c,1}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}}\right] \\
=& \underbrace{\hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\frac{(1-\pi)X^c_{n^-}}{(1-\pi)X^c_{n^-}+\pi Y^c_{n^+}} - \frac{(1-\pi)X^c_{n^-}}{(1-\pi)\mu_{c,1}+\pi Y^c_{n^+}}\right]}_{(a)} \\
&+ \underbrace{\hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\frac{(1-\pi)X^c_{n^-}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}} - \frac{(1-\pi)\mu_{c,1}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}}\right]}_{(b)} \\
&+ \underbrace{\hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\frac{(1-\pi)X^c_{n^-}}{(1-\pi)\mu_{c,1}+\pi Y^c_{n^+}} - \frac{(1-\pi)X^c_{n^-}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}}\right]}_{(c)} \\
&+ \underbrace{\hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\frac{(1-\pi)\mu_{c,1}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}}\right] - \mathbb{E}_{c\sim h_{\boldsymbol{w}}(\mathcal{S}^+)}\left[\frac{(1-\pi)\mu_{c,1}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}}\right]}_{(d)}.
\end{aligned}
\tag{25}
$$

Notice that $\boldsymbol{z}^+$ is randomly sampled from $\mathcal{S}^+$, thus we have $(d)=0$, and we only need to focus on the first three terms.

To solve (a), for any threshold c, we consider a continuous function $s(;c)$:

$$
s(x;c) = \frac{(1-\pi)X^c_{n^-}}{(1-\pi)x+\pi Y^c_{n^+}}.
$$

With the delta method [67], when $n^-\to\infty$ we have

$$
s\left(X^c_{n^-};c\right) - s\left(\mu_{c,1};c\right) \rightsquigarrow \mathcal{N}(0,(s'(\mu_{c,1}))^2\kappa^2_{c,1}/n^-). \tag{26}
$$

Notice that the above result holds for all $\boldsymbol{z}$ and c. Consider a random variable $\boldsymbol{X}_{n^-}\in\mathbb{R}^{Mn^+}$ whose elements are $X^c_{n^-}$ w.r.t. all c, and a function $Avg:\mathbb{R}^{Mn^+}\mapsto\mathbb{R}$ outputs the average of the input. Obviously, the gradient of Avg at any point is a scaled unit vector. Therefore, according to the delta method, we have the following convergence:

$$
(a) = \hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[s\left(X^c_{n^-};c\right) - s\left(\mu_{c,1};c\right)\right] \rightsquigarrow \mathcal{N}(0,\|\Sigma\|_1/(Mn^+)^2), \tag{27}
$$

where $\Sigma = Var[\boldsymbol{X}_{n^-}]$. Denote

$$
t_1^2 = \hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[(s'(\mu_{c,1}))^2\kappa^2_{c,1}\right],\ \ \kappa_1^2 = \hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\kappa^2_{c,1}\right],\ \ \kappa_2^2 = \hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\kappa^2_{c,2}\right]
$$

then we have

$$
(t_1^2/n^-)/(Mn^+) \le \|\Sigma\|_1/(Mn^+)^2 \le t_1^2/n^-. \tag{28}
$$

According to the mean value theorem, there exists a positive constant ζ_a such that

$$
\|\Sigma\|_1/(Mn^+)^2 = \zeta_a\cdot\kappa_1^2/n^-, \tag{29}
$$

and (27) can be rewritten as

$$
(a) = \hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[s\left(X^c_{n^-};c\right) - s\left(\mu_{c,1};c\right)\right] \rightsquigarrow \mathcal{N}(0,\zeta_a\cdot\kappa_1^2/n^-). \tag{30}
$$

Similarly, there exists $\zeta_b,\zeta_c>0$, such that

$$
\begin{aligned}
(b) &\rightsquigarrow \mathcal{N}\left(0,\ \zeta_b\cdot\kappa_1^2/n^-\right), \\
(c) &\rightsquigarrow \mathcal{N}\left(0,\ \zeta_c\cdot\kappa_2^2/n^+\right).
\end{aligned}
\tag{31}
$$

To sum up, when $\hat{\mathbb{E}}_{c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\kappa_{c,1}^2\right]/n^- \to 0$ and $\hat{\mathbb{E}}_{c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\kappa_{c,2}^2\right]/n^+ \to 0$, we have

$$(a) + (b) + (c) \xrightarrow{P} 0, \tag{32}$$

where $\xrightarrow{P}$ refers to convergence in probability.

As for $\widehat{AP}^{\downarrow}$, the difference can be decomposed into

$$\begin{aligned}
&\hat{\mathbb{E}}_{\boldsymbol{z}\subseteq\mathcal{S}}\left[\widehat{AP}^{\downarrow}(\boldsymbol{w};\boldsymbol{z})\right] - \widehat{\text{AUPRC}}^{\downarrow}(\boldsymbol{w};\mathcal{S}) \\
=& \hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\frac{(1-\pi_0)X_{n^-}^c}{(1-\pi_0)X_{n^-}^c + \pi_0 Y_{n^+}^c}\right] - \mathbb{E}_{c\sim h_{\boldsymbol{w}}(\mathcal{S}^+)}\left[\frac{(1-\pi)\mu_{c,1}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}}\right] \\
=& \underbrace{\hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\frac{(1-\pi_0)X_{n^-}^c}{(1-\pi_0)X_{n^-}^c + \pi_0 Y_{n^+}^c}\right] - \mathbb{E}_{c\sim h_{\boldsymbol{w}}(\mathcal{S}^+)}\left[\frac{(1-\pi_0)\mu_{c,1}}{(1-\pi_0)\mu_{c,1}+\pi_0\mu_{c,2}}\right]}_{(e)} \\
&+ \underbrace{\mathbb{E}_{c\sim h_{\boldsymbol{w}}(\mathcal{S}^+)}\left[\frac{(1-\pi_0)\mu_{c,1}}{(1-\pi_0)\mu_{c,1}+\pi_0\mu_{c,2}}\right] - \mathbb{E}_{c\sim h_{\boldsymbol{w}}(\mathcal{S}^+)}\left[\frac{(1-\pi)\mu_{c,1}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}}\right]}_{(f)}.
\end{aligned} \tag{33}$$

Similarly, we have $(e) \xrightarrow{P} 0$. As for (f), we have

$$(f) = (\pi_0 - \pi)\cdot \mathbb{E}_{c\sim h_{\boldsymbol{w}}(\mathcal{S}^+)}\left[\frac{\mu_{c,2}/(1-\pi)}{(1-\pi_0)\mu_{c,1}+\pi_0\mu_{c,2}}\cdot\frac{(1-\pi)\mu_{c,1}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}}\right]. \tag{34}$$

According to the mean value theorem, there exists

$$H \in \left[\min_{\boldsymbol{x}\sim\mathcal{S}^+}\frac{\mu_{c,2}/(1-\pi)}{(1-\pi_0)\mu_{c,1}+\pi_0\mu_{c,2}}, \max_{\boldsymbol{x}\sim\mathcal{S}^+}\frac{\mu_{c,2}/(1-\pi)}{(1-\pi_0)\mu_{c,1}+\pi_0\mu_{c,2}}\right], \tag{35}$$

such that

$$(f) = (\pi_0 - \pi)H\cdot \mathbb{E}_{c\sim h_{\boldsymbol{w}}(\mathcal{S}^+)}\left[\frac{(1-\pi)\mu_{c,1}}{(1-\pi)\mu_{c,1}+\pi\mu_{c,2}}\right]. \tag{36}$$

□

Proposition 4. *For any $0<\delta<1$, at least with probability of $1-\delta$, we have*

$$\left|\hat{\mathbb{E}}_{\boldsymbol{z}\subseteq\mathcal{S}}[\hat{f}(\boldsymbol{w};\boldsymbol{z})] - \widehat{AUPRC}^{\downarrow}(\boldsymbol{w};\mathcal{S})\right| = \mathcal{O}\left(\sqrt{\frac{\log(6n^+/\delta)}{n^+}} + 2\sqrt{\frac{\log(6n^+/\delta)}{n^-}}\right).$$

Proof. Denote $X_{n^-}^c$, $Y_{n^+}^c$, (a), (b), (c) as in the proof of Prop. 2. Next, we focus on the term (a). Under the assumption of Prop. 2, $X_{n^-}^c$ can be viewed as an average of i.i.d. variables, thus according to Hoeffding's inequality, for any $\epsilon>0$ we have

$$\mathbb{P}\left(|X_{n^-}^c - \mu_{c,1}| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^-\epsilon^2}{B_{\ell_1}^2}\right).$$

Therefore, for any c, $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\left| \frac{(1-\pi)X^c_{n^-}}{(1-\pi)X^c_{n^-} + \pi Y^c_{n^+}} - \frac{(1-\pi)X^c_{n^-}}{(1-\pi)\mu_{c,1} + \pi Y^c_{n^+}} \right|$$

$$\leq \left| \frac{(1-\pi)^2 X^c_{n^-}}{\left((1-\pi)X^c_{n^-} + \pi Y^c_{n^+}\right)\left((1-\pi)\mu_{c,1} + \pi Y^c_{n^+}\right)} \right| \cdot |X^c_{n^-} - \mu_{c,1}|$$

$$\leq \frac{1}{\mu_{c,1}} \cdot |X^c_{n^-} - \mu_{c,1}|$$

$$\leq \sqrt{\frac{B^2_{\ell_1} \log \frac{2}{\delta}}{2\mu^2_{c,1} n^-}}$$

$$\leq \sqrt{\frac{B^2_{\ell_1} \log \frac{2}{\delta}}{2\mu^2_1 n^-}},$$

where $\mu_1 = \inf_c \mu_{c,1}$. If we further assume that $X^c_{n^-}$ is independent w.r.t. different c, then considering all positive c, with probability at least $1 - \delta/3$, we have

$$|(a)| \leq \sqrt{\frac{B^2_{\ell_1} \log \frac{6n^+}{\delta}}{2\mu^2_1 n^-}} = \mathcal{O}\left(\sqrt{\frac{\log(6n^+/\delta)}{n^-}}\right),$$

and similarly

$$|(b)| = \mathcal{O}\left(\frac{\log(6n^+/\delta)}{n^-}\right), \quad |(c)| = \mathcal{O}\left(\sqrt{\frac{\log(6n^+/\delta)}{n^+}}\right).$$

To sum up, with probability at least $1 - \delta$ we have

$$\left| \hat{\mathbb{E}}_{\boldsymbol{z} \subseteq \mathcal{S}}[\hat{f}(\boldsymbol{w}; \boldsymbol{z})] - \widehat{\text{AUPRC}}^{\downarrow}(\boldsymbol{w}; \mathcal{S}) \right| \leq |(a)| + |(b)| + |(c)| = \mathcal{O}\left(\sqrt{\frac{\log(6n^+/\delta)}{n^+}} + 2\sqrt{\frac{\log(6n^+/\delta)}{n^-}}\right).$$

□

B.2 Details of Simulation Experiments

Following [8], assume that scores are i.i.d. random variables drawn from one of three distributions: binormal, bibeta and uniform offset. The hyperparameters choosen and the corresponding probability density functions are shown in Fig. 3. For each kind of distribution, we draw $100,000$ score points, where the ratio of positive points $\pi = 0.1$. Afterward, under different sampling-rate $\pi_0 \in \{0.01, 0.02, 0.03, 0.1, 0.2\}$, we repeatly sample several points from the above-mentioned points for 500 times, and report the corresponding mean value and standard derivation. The results are shown in Fig. 4 and Fig. 5, which are consistent with our conclusions in the main paper.

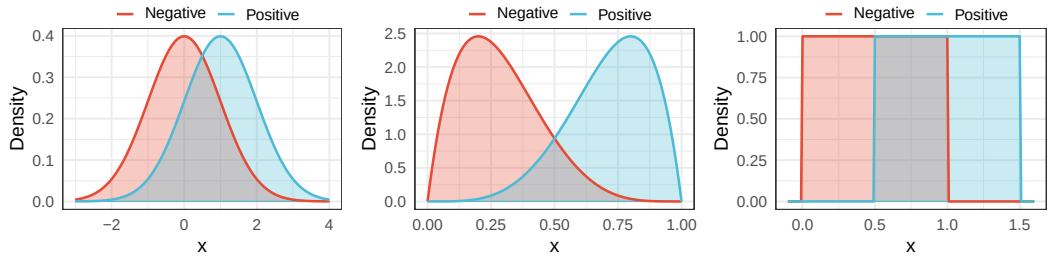
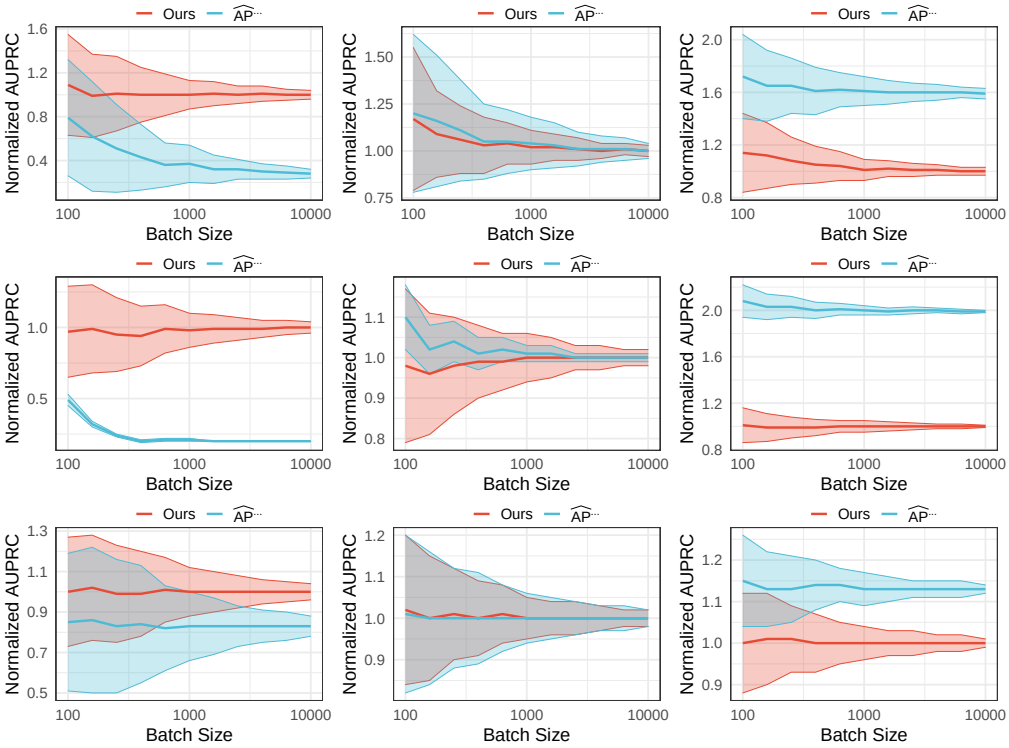
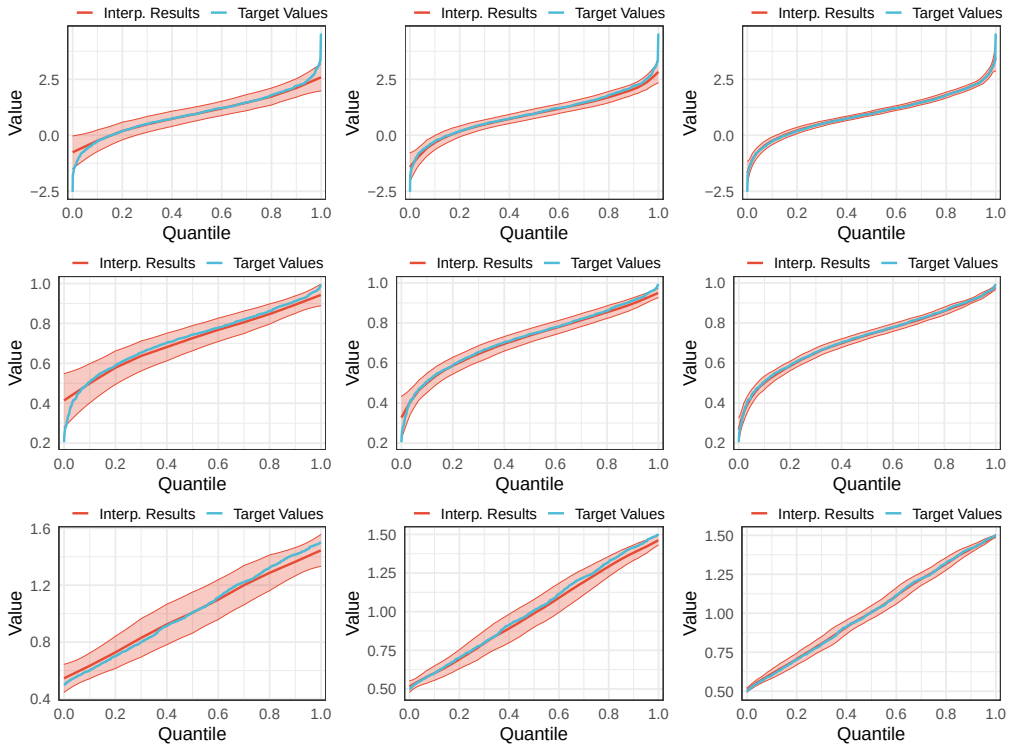


Figure 3: Probability density functions of positive scores s^+ and negative scores s^-. Left: binormal ($s^- \sim \mathcal{N}(0,1), s^+ \sim \mathcal{N}(1,1)$). Middle: bibeta ($s^- \sim \mathcal{B}(2,5), s^+ \sim \mathcal{B}(5,2)$). Right: offset uniform ($s^- \sim \mathcal{U}(0,1), s^+ \sim \mathcal{U}(0.5,1.5)$).



(a) $\pi_0 = 0.02$ (b) $\pi_0 = 0.1$ (c) $\pi_0 = 0.2$

Figure 4: Empirical analysis of estimation errors on simulation data. The score distributions of each row are binormal, bibeta and uniform offset in turn.



(a) $\pi_0 = 0.01$ (b) $\pi_0 = 0.03$ (c) $\pi_0 = 0.1$

Figure 5: Empirical analysis of interpolation errors on simulation data. The score distributions of each row are binormal, bibeta and uniform offset in turn.

C Proofs on Generalization via Stability

C.1 Generalization by List Model Stability

Theorem 1 (Generalization via Model Stability). *Let a stochastic algorithm A be listwise on-average model (ϵ^+, ϵ^-)-stable and Asmp. 1 holds. Then we have*

$$\mathbb{E}_{\mathcal{S},A}\left[F(A(\mathcal{S})) - F(A(\mathcal{S});\mathcal{S})\right] \leq G(n^+\epsilon^+ + n^-\epsilon^-). \tag{12}$$

Proof. Let $\mathcal{S}, \widetilde{\mathcal{S}}$ be defined as in Def. 1. First of all, according to the symmetry, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S},A}[F(A(\mathcal{S})) - F(A(\mathcal{S});\mathcal{S})] \\
=&\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},A}[F(A(\mathcal{S});\widetilde{\mathcal{S}}) - F(A(\mathcal{S});\mathcal{S})] \\
=&\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},A}[F(A(\widetilde{\mathcal{S}});\mathcal{S}) - F(A(\mathcal{S});\mathcal{S})] \\
=&\frac{1}{M}\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},A}\left[\sum_{\boldsymbol{z}\sim\mathcal{S}}\left(\hat{f}(A(\widetilde{\mathcal{S}});\boldsymbol{z}) - \hat{f}(A(\mathcal{S});\boldsymbol{z})\right)\right] \\
=&\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},\boldsymbol{z},\tilde{\boldsymbol{z}},A}\left[\hat{f}(A((\mathcal{S}-\boldsymbol{z})\cup\tilde{\boldsymbol{z}});\boldsymbol{z}) - \hat{f}(A(\mathcal{S});\boldsymbol{z})\right],
\end{aligned} \tag{37}$$

where $\boldsymbol{z} \subseteq \mathcal{S}$ and $\tilde{\boldsymbol{z}} \subseteq \widetilde{\mathcal{S}}$. Then, since the gradients of $\hat{f}$ is upper bounded by G, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},\boldsymbol{z},\tilde{\boldsymbol{z}},A}\left[\left|\hat{f}(A((\mathcal{S}-\boldsymbol{z})\cup\tilde{\boldsymbol{z}});\boldsymbol{z})-\hat{f}(A(\mathcal{S});\boldsymbol{z})\right|\right]\\
\leq& G\cdot\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},\boldsymbol{z},\tilde{\boldsymbol{z}},A}\left[\|A((\mathcal{S}-\boldsymbol{z})\cup\tilde{\boldsymbol{z}})-A(\mathcal{S})\|_2\right]\\
\leq& Gn^+\cdot\mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\left[\left\|A(S^{(i)})-A(\mathcal{S})\right\|_2|y_i=1\right]\\
&+Gn^-\cdot\mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\left[\left\|A(S^{(i)})-A(\mathcal{S})\right\|_2|y_i=-1\right]\\
\leq&\frac{Gn^+}{N^+}\mathbb{E}_{\mathcal{S},\widehat{\mathcal{S}},A}\left[\sum_{y_i=1}\left\|A(S^{(i)})-A(\mathcal{S})\right\|_2\right]+\frac{Gn^-}{N^-}\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},A}\left[\sum_{y_i=-1}\left\|A(S^{(i)})-A(\mathcal{S})\right\|_2\right].
\end{aligned}\tag{38}$$

The proof is completed by denoting

$$\begin{aligned}
\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},A}\left[\frac{1}{N^+}\sum_{y_i=1}\left\|A(S^{(i)})-A(\mathcal{S})\right\|_2\right]&=\epsilon^+,\\
\mathbb{E}_{\mathcal{S},\widetilde{\mathcal{S}},A}\left[\frac{1}{N^-}\sum_{y_i=-1}\left\|A(S^{(i)})-A(\mathcal{S})\right\|_2\right]&=\epsilon^-.
\end{aligned}\tag{39}$$

□

C.2 List Model Stability of SOPRC

Lemma 1. *Let $\mathcal{S},\widetilde{\mathcal{S}},\mathcal{S}^{(i)}$ be constructed as Def. 1 and Asmp. 1, 2, 3 hold. Let $\{\boldsymbol{w}_t\}_t$ and $\{\boldsymbol{w}_t^{(i)}\}_t$ be produced by Alg. 1 with $\mathcal{S}$ and $\mathcal{S}^{(i)}$, respectively. Denote $L=\max\{L_w,L_v/n^+,C_\phi B,G/2,B'_\ell\}$, $\boldsymbol{m}_t^{(i)}=\begin{bmatrix}\|\boldsymbol{w}_t-\boldsymbol{w}_t^{(i)}\|_2 & \|\boldsymbol{v}_t-\boldsymbol{v}_t^{(i)}\|_2 & 1\end{bmatrix}^\top$, $\boldsymbol{m}_t^+=\frac{1}{N^+}\sum_{y_i=1}\mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}_t^{(i)}\right]$, $\boldsymbol{m}_t^-=\frac{1}{N^+}\sum_{y_i=-1}\mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}_t^{(i)}\right]$. Then for all $t\in[T]$, by setting $\beta_t\leq 2C_\phi B/n^+$, we have*

$$\boldsymbol{m}_{t+1}^+\leq\left(\boldsymbol{I}_3+\boldsymbol{R}_t^+\right)\cdot\boldsymbol{m}_t^+,\qquad \boldsymbol{m}_{t+1}^-\leq\left(\boldsymbol{I}_3+\boldsymbol{R}_t^-\right)\cdot\boldsymbol{m}_t^-,\tag{13}$$

where $\boldsymbol{I}_3$ is the 3×3 identity matrix and

$$R_t^+=\begin{bmatrix}2L\eta_t & \frac{L(1-\beta_t)\eta_t}{N^+} & \frac{L\eta_t}{N^+}\\ L\beta_t & 0 & \frac{1}{N^+}\\ 0 & 0 & 0\end{bmatrix},R_t^-=\begin{bmatrix}2L\eta_t & \frac{L_v(1-\beta_t)\eta_t}{N^+} & \frac{L\eta_t\cdot n^+}{N^-}\\ L\beta_t & 0 & 0\\ 0 & 0 & 0\end{bmatrix}.\tag{14}$$

Proof. For all $i=1,\cdots,N$, let $\{\boldsymbol{w}_t\}_t,\{\boldsymbol{v}_t\}_t,\{\boldsymbol{z}_{i_t}\}_t$ and $\{\boldsymbol{w}'_t\}_t,\{\boldsymbol{v}'_t\}_t,\{\boldsymbol{z}'_{i_t}\}_t$ be produced by Alg. 1 based on $\mathcal{S}$ and $\mathcal{S}^{(i)}$, respectively. According to the update rules, for all $t=1,\cdots,T$, we have

$$\begin{aligned}
\|\boldsymbol{w}'_{t+1}-\boldsymbol{w}_{t+1}\|_2&=\|\boldsymbol{w}'_t-\boldsymbol{w}_t-\eta_t\left(\nabla f(\boldsymbol{w}'_t;\boldsymbol{z}'_{i_t},\boldsymbol{v}'_{t+1})-\nabla f(\boldsymbol{w}_t;\boldsymbol{z}_{i_t},\boldsymbol{v}_{t+1})\right)\|_2\\
&\leq\|\boldsymbol{w}'_t-\boldsymbol{w}_t\|_2+\eta_t\left\|\nabla f(\boldsymbol{w}'_t;\boldsymbol{z}'_{i_t},\boldsymbol{v}'_{t+1})-\nabla f(\boldsymbol{w}_t;\boldsymbol{z}_{i_t},\boldsymbol{v}_{t+1})\right\|_2.
\end{aligned}\tag{40}$$

We discuss the second term on the right side in the following three cases.

Case 1: $(\boldsymbol{x}_i,y_i)\notin\boldsymbol{z}_{i_t}$. In this case, we have $\boldsymbol{z}_{i_t}=\boldsymbol{z}'_{i_t}$. According to the smoothness of f and Lem. 4, we have

$$\begin{aligned}
&\left\|\nabla f(\boldsymbol{w}'_t;\boldsymbol{z}'_{i_t},\boldsymbol{v}'_{t+1})-\nabla f(\boldsymbol{w}_t;\boldsymbol{z}_{i_t},\boldsymbol{v}_{t+1})\right\|_2\\
\leq&\left\|\nabla f(\boldsymbol{w}'_t;\boldsymbol{z}_{i_t},\boldsymbol{v}'_{t+1})-\nabla f(\boldsymbol{w}_t;\boldsymbol{z}_{i_t},\boldsymbol{v}'_{t+1})\right\|\\
&+\left\|\nabla f(\boldsymbol{w}_t;\boldsymbol{z}_{i_t},\boldsymbol{v}'_{t+1})-\nabla f(\boldsymbol{w}_t;\boldsymbol{z}_{i_t},\boldsymbol{v}_{t+1})\right\|_2\\
\leq&L_w\|\boldsymbol{w}'_t-\boldsymbol{w}_t\|_2+L_v\|\boldsymbol{v}'_{t+1}-\boldsymbol{v}_{t+1}\|_2/N^+,
\end{aligned}\tag{41}$$

and

$$\begin{aligned}
&\|\boldsymbol{v}'_{t+1}-\boldsymbol{v}_{t+1}\|_2\\
\leq&(1-\beta_t)\|\boldsymbol{v}'_t-\boldsymbol{v}_t\|_2+\beta_t\|\phi\left(h_{\boldsymbol{w}'_t}(\boldsymbol{z}_{i_t}^+)\right)-\phi\left(h_{\boldsymbol{w}_t}(\boldsymbol{z}_{i_t}^+)\right)\|_2\\
\leq&(1-\beta_t)\|\boldsymbol{v}'_t-\boldsymbol{v}_t\|_2+C_\phi C_h\beta_t\|\boldsymbol{w}'_t-\boldsymbol{w}_t\|_2\\
\leq&(1-\beta_t)\|\boldsymbol{v}'_t-\boldsymbol{v}_t\|_2+\|\boldsymbol{w}'_t-\boldsymbol{w}_t\|_2/n^+,
\end{aligned}\tag{42}$$

where the last inequation is due to $\beta_t \leq \frac{C_\phi B}{n^+}$.

Combining Eq. (40), Eq. (41) and Eq. (42), we have

$$\begin{aligned}\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_{t+1}\|_2 \leq &(1 + L_w\eta_t + L_v\eta_t/n^+)\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2 \\ &+ L_v(1-\beta_t)\eta_t\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2/N^+\end{aligned} \tag{43}$$

Case 2: $(\boldsymbol{x}_i, y_i) \in \boldsymbol{z}_{i_t}$ and $y_i = 1$. In this case, we have $\boldsymbol{z}^-_{i_t} = \boldsymbol{z}'^-_{i_t}$, thus

$$\begin{aligned}&\left\|\nabla f(\boldsymbol{w}'_t; \boldsymbol{z}'_{i_t}, \boldsymbol{v}'_{t+1}) - \nabla f(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}_{t+1})\right\|_2 \\ \leq &G/n^+ + L_w\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2 + L_v\|\boldsymbol{v}'_{t+1} - \boldsymbol{v}_{t+1}\|_2/N^+,\end{aligned} \tag{44}$$

and

$$\begin{aligned}&\|\boldsymbol{v}'_{t+1} - \boldsymbol{v}_{t+1}\|_2 \\ \leq &(1-\beta_t)\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2 + \beta_t\|\phi\left(h_{\boldsymbol{w}'_t}(\boldsymbol{z}'^+_{i_t})\right) - \phi\left(h_{\boldsymbol{w}_t}(\boldsymbol{z}'^+_{i_t})\right)\|_2 \\ &+ \beta_t\|\phi\left(h_{\boldsymbol{w}_t}(\boldsymbol{z}'^+_{i_t})\right) - \phi\left(h_{\boldsymbol{w}_t}(\boldsymbol{z}^+_{i_t})\right)\|_2 \\ \leq &(1-\beta_t)\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2 + \|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2/n^+ + 2C_\phi B\beta_t \\ \leq &(1-\beta_t)\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2 + \|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2/n^+ + 1/n^+,\end{aligned} \tag{45}$$

where the last inequality is due to $\beta_t \leq 2C_\phi B/n^+$.

By combining Eq. (44) and Eq. (45), we have

$$\begin{aligned}\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_{t+1}\|_2 \leq &(1 + L_w\eta_t + L_v\eta_t/n^+)\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2 \\ &+ L_v(1-\beta_t)\eta_t\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2/N^+ + (G + L_v/N^+)\eta_t/n^+.\end{aligned} \tag{46}$$

Case 3: $(\boldsymbol{x}_i, y_i) \in \boldsymbol{z}_{i_t}$ and $y_i = -1$. In this case, we have $\boldsymbol{z}^+_{i_t} = \boldsymbol{z}'^+_{i_t}$, thus

$$\begin{aligned}&\left\|\nabla f(\boldsymbol{w}'_t; \boldsymbol{z}'_{i_t}, \boldsymbol{v}'_{t+1}) - \nabla f(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}_{t+1})\right\|_2 \\ \leq &\left\|\nabla f(\boldsymbol{w}'_t; \boldsymbol{z}'_{i_t}, \boldsymbol{v}'_{t+1}) - \nabla f(\boldsymbol{w}'_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}'_{t+1})\right\|_2 \\ &+ \left\|\nabla f(\boldsymbol{w}'_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}'_{t+1}) - \nabla f(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}'_{t+1})\right\|_2 \\ &+ \left\|\nabla f(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}'_{t+1}) - \nabla f(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}_{t+1})\right\|_2 \\ \leq &B'_\ell \cdot n^+/n^- + L_w\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2 + L_v\|\boldsymbol{v}'_{t+1} - \boldsymbol{v}_{t+1}\|_2/N^+,\end{aligned} \tag{47}$$

where $B'_\ell = (1-\pi)B_\ell/\pi$. Similar to case 1, Eq. (42) still holds. Combining it with Eq. (47), we have

$$\begin{aligned}\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_{t+1}\|_2 \leq &(1 + L_w\eta_t + L_v\eta_t/n^+)\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2 \\ &+ L_v\eta_t\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2/N^+ + B'_\ell \cdot n^+/n^-\end{aligned} \tag{48}$$

Next, we consider the expectation of $\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_{t+1}\|_2$ and $\|\boldsymbol{v}'_{t+1} - \boldsymbol{v}_{t+1}\|_2$ taking on A. Note that

$$\begin{aligned}&\mathbb{P}(\text{case 1}|y_i = 1) = (N^+ - n^+)/N^+, \qquad \mathbb{P}(\text{case 3}|y_i = 1) = n^+/N^+, \\ &\mathbb{P}(\text{case 1}|y_i = -1) = (N^- - n^-)/N^-, \quad \mathbb{P}(\text{case 2}|y_i = -1) = n^-/N^-,\end{aligned} \tag{49}$$

For $y_i = 1$, we have

$$\begin{aligned}\mathbb{E}_A\left[\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_{t+1}\|_2\right] \leq &(1 + L_w\eta_t + L_v\eta_t/n^+)\mathbb{E}_A\left[\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2\right] \\ &+ L_v(1-\beta_t)\eta_t\mathbb{E}_A\left[\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2\right]/N^+ + (G + L_v/N^+)/N^+, \\ \mathbb{E}_A\left[\|\boldsymbol{v}'_{t+1} - \boldsymbol{v}_{t+1}\|_2\right] \leq &C_\phi B\beta_t\mathbb{E}_A\left[\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2\right] + \mathbb{E}_A\left[\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2\right] + 1/N^+.\end{aligned} \tag{50}$$

By setting $L \geq \max\{L_w, L_v/n^+, C_\phi B, G/2\}$ and $N^+ > 2n^+$, the above result can be rewritten as

$$\begin{aligned}&\begin{bmatrix}\mathbb{E}_A\left[\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_{t+1}\|_2\right] \\ \mathbb{E}_A\left[\|\boldsymbol{v}'_{t+1} - \boldsymbol{v}_{t+1}\|_2\right] \\ 1\end{bmatrix} \\ \leq &\begin{bmatrix}1 + 2L\eta_t & L(1-\beta_t)\eta_t/N^+ & L\eta_t/N^+ \\ L\beta_t & 1 & 1/N^+ \\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\mathbb{E}_A\left[\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2\right] \\ \mathbb{E}_A\left[\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2\right] \\ 1\end{bmatrix}\end{aligned} \tag{51}$$

Similarly, for $y_i = -1$, we have

$$\begin{aligned}\mathbb{E}_A\left[\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_{t+1}\|_2\right] \leq & (1 + L_w\eta_t + L_v\eta_t/n^+)\mathbb{E}_A\left[\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2\right] \\ & + L_v(1-\beta_t)\eta_t\mathbb{E}_A\left[\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2\right]/N^+ + B'_\ell n^+/N^-, \\ \mathbb{E}_A\left[\|\boldsymbol{v}'_{t+1} - \boldsymbol{v}_{t+1}\|_2\right] \leq & C_\phi B\beta_t\mathbb{E}_A\left[\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2\right] + \mathbb{E}_A\left[\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2\right].\end{aligned} \tag{52}$$

By setting $L \geq \max\{L_w, L_v/n^+, C_\phi B, G, B'_\ell\}$, we have

$$\begin{aligned}&\begin{bmatrix}\mathbb{E}_A\left[\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_{t+1}\|_2\right] \\ \mathbb{E}_A\left[\|\boldsymbol{v}'_{t+1} - \boldsymbol{v}_{t+1}\|_2\right] \\ 1\end{bmatrix} \\ \leq &\begin{bmatrix}1+2L\eta_t & L_v(1-\beta_t)\eta_t/N^- & L\eta_t\cdot n^+/N^- \\ L\beta_t & 1 & 0 \\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\mathbb{E}_A\left[\|\boldsymbol{w}'_t - \boldsymbol{w}_t\|_2\right] \\ \mathbb{E}_A\left[\|\boldsymbol{v}'_t - \boldsymbol{v}_t\|_2\right] \\ 1\end{bmatrix}.\end{aligned} \tag{53}$$

□

Lemma 2. *Let $\boldsymbol{z}$, $\boldsymbol{z}' \in \mathcal{Z}^n$ be two datasets that differ by exact one example and Asmp. 1 holds. For all $\boldsymbol{w} \in \Omega, \boldsymbol{v} \in \mathbb{R}^{N^+}$, the gradient difference of w.r.t. $\boldsymbol{z}$ and $\boldsymbol{z}'$ are upper bounded:*

(a) If $\boldsymbol{z}, \boldsymbol{z}'$ differ by a negative example, then

$$\|\nabla f(\boldsymbol{w}; \boldsymbol{z}'_{i_t}, v) - \nabla f(\boldsymbol{w}; \boldsymbol{z}_{i_t}, v)\|_2 \leq \frac{(1-\pi)B_\ell|\boldsymbol{z}^+|}{\pi|\boldsymbol{z}^-|} \tag{54}$$

(b) If $\boldsymbol{z}, \boldsymbol{z}'$ differ by a positive example, then

$$\|\nabla f(\boldsymbol{w}; \boldsymbol{z}'_{i_t}, v) - \nabla f(\boldsymbol{w}; \boldsymbol{z}_{i_t}, v)\|_2 \leq G/|\boldsymbol{z}^+| \tag{55}$$

Proof. The proof of (b) is obvious. For (a), let the different examples in $\boldsymbol{z}$ and $\boldsymbol{z}'$ be $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}'$.

Denote $c(\boldsymbol{x}; \boldsymbol{v}, \ell) = \hat{\mathbb{E}}_{v\sim\boldsymbol{v}}[\ell(h_{\boldsymbol{w}}(\boldsymbol{x}) - v))]$, and rewrite f as

$$f(\boldsymbol{w}; \boldsymbol{z}, \boldsymbol{v}) = \hat{\mathbb{E}}_{\boldsymbol{x}\in\boldsymbol{z}^+}\left[\sigma\left(\frac{1-\pi}{\pi}\cdot c(\boldsymbol{x}; h_{\boldsymbol{w}}(\boldsymbol{z}^-), \ell_1)/c(\boldsymbol{x}; v, \ell_2)\right)\right]. \tag{56}$$

$$\begin{aligned}&\|\nabla f(\boldsymbol{w}; \boldsymbol{z}'_{i_t}, v) - \nabla f(\boldsymbol{w}; \boldsymbol{z}_{i_t}, v)\|_2 \\ \leq & \hat{\mathbb{E}}_{\boldsymbol{x}\in\boldsymbol{z}^+}\Big[\|\nabla\sigma\left((1-\pi)/\pi\cdot c(\boldsymbol{x}; h_{\boldsymbol{w}}(\boldsymbol{z}^-), \ell_1)/c(\boldsymbol{x}; \boldsymbol{v}, \ell_2)\right) \\ & - \nabla\sigma\left((1-\pi)/\pi\cdot c(\boldsymbol{x}; h_{\boldsymbol{w}}(\boldsymbol{z}'^-), \ell_1)/c(\boldsymbol{x}; \boldsymbol{v}, \ell_2)\right)\|_2\Big] \\ \leq & \hat{\mathbb{E}}_{\boldsymbol{x}\in\boldsymbol{z}^+}\Big[\|(1-\pi)/\pi\cdot c(\boldsymbol{x}; h_{\boldsymbol{w}}(\boldsymbol{z}^-), \ell_1)/c(\boldsymbol{x}; \boldsymbol{v}, \ell_2) \\ & - (1-\pi)/\pi\cdot c(\boldsymbol{x}; h_{\boldsymbol{w}}(\boldsymbol{z}'^-), \ell_1)/c(\boldsymbol{x}; \boldsymbol{v}, \ell_2)\|_2\Big] \\ \leq & \frac{(1-\pi)|\boldsymbol{z}^+|}{\pi}\hat{\mathbb{E}}_{\boldsymbol{x}\in\boldsymbol{z}^+}\left[\|c(\boldsymbol{x}; h_{\boldsymbol{w}}(\boldsymbol{z}^-), \ell_1) - c(\boldsymbol{x}; h_{\boldsymbol{w}}(\boldsymbol{z}'^-), \ell_1)\|_2\right] \\ \leq & \frac{(1-\pi)B_\ell|\boldsymbol{z}^+|}{\pi|\boldsymbol{z}^-|}.\end{aligned} \tag{57}$$

□

Lemma 3. *For two datasets $\mathcal{S}, \mathcal{S}^{(i)}$ that differ by the i-th element, let $\{\boldsymbol{w}_t\}_t, \{\boldsymbol{v}_t\}_t$ and $\{\boldsymbol{w}'_t\}_t, \{\boldsymbol{v}'_t\}_t$ be produced by Alg. 1 with $\mathcal{S}$ and $\mathcal{S}^{(i)}$, respectively. If $y_i = 1$, then for all $t_0 \in [N^+/n^+]$ we have*

$$\mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}^{(i)}_{T+1}\right] \leq \frac{t_0 n^+}{N^+}\begin{bmatrix} D(\Omega) & B & 1\end{bmatrix}^\top + \mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}^{(i)}_{T+1} \mid i \notin I_{t_0}(A)\right], \tag{58}$$

where $D(\Omega)$ is the diameter of the hypothesis space, $I_t(A) := \{\boldsymbol{i}_1, \cdots, \boldsymbol{i}_t\}$ is the set of indices selected by A at the first t-th iterations. Similarly, when $y_i = -1$, for all $t_0 \in [N^-/n^-]$, we have

$$\mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}^{(i)}_{T+1}\right] \leq \frac{t_0 n^-}{N^-}\begin{bmatrix} D(\Omega) & B & 1\end{bmatrix}^\top + \mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}^{(i)}_{T+1} \mid i \notin I_{t_0}(A)\right], \tag{59}$$

Proof. Here we only need to prove the case $y_i = 1$. According to the law of total probability, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}_{T+1}^{(i)}\right] \leq & \mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}_{T+1}^{(i)} \mid i \in I_{t_0}(A)\right]\mathbb{P}\left(i \in I_{t_0}(A)\right) \\ & + \mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}_{t_0}^{(i)} \mid i \notin I_{T+1}(A)\right]\mathbb{P}\left(i \notin I_{t_0}(A)\right).\end{aligned} \tag{60}$$

Notice that

$$\mathbb{P}\left(i \in I_{t_0}(A)\right) \leq \sum_{t=1}^{t_0}\mathbb{P}\left(i \in \boldsymbol{i}_t\right) = \frac{t_0 n^+}{N^+}, \tag{61}$$

the proof is complete by bounding the first term on the right side of Eq. (60). □

Theorem 2. *Let $\lambda = LC_\eta(1+\sqrt{1-\beta^2+\beta})$, and assumptions in Lem. 1 hold. By setting $\eta_t \leq \frac{C_\eta}{t}$, $\beta_t = \beta \asymp 1/n^+$ and $T \leq N^+$, Alg. 1 is list on-average model stable with*

$$\epsilon^+ = \mathcal{O}\left(\frac{(Tn^+)^{\frac{\lambda}{\lambda+1}}}{N^+}\right), \epsilon^- = \mathcal{O}\left(\frac{(Tn^-)^{\frac{\lambda}{\lambda+1}}}{N^-}\right). \tag{15}$$

Proof. According to Lem. 1 and Lem. 3, we have

$$\begin{aligned}\frac{1}{N^+}\sum_{y_i=1}\mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}_{T+1}^{(i)}\right] \leq & \frac{t_0 n^+}{N^+}\begin{bmatrix} D(\Omega) & B & 1 \end{bmatrix}^\top \\ & + \prod_{t=t_0}^{T}(\boldsymbol{I}_3 + \boldsymbol{R}_t^+)\cdot\frac{1}{N^+}\sum_{y_i=1}\mathbb{E}_{\mathcal{S},A}\left[\boldsymbol{m}_{t_0}^{(i)} \middle| i \notin I_{t_0}(A)\right] \\ = & \frac{t_0 n^+}{N^+}\begin{bmatrix} D(\Omega) & B & 1 \end{bmatrix}^\top + \prod_{t=t_0}^{T}(\boldsymbol{I}_3 + \boldsymbol{R}_t^+)\cdot\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^\top,\end{aligned} \tag{62}$$

where the last equation is due to the fact that for all $t \leq t_0$, $\boldsymbol{z}_{i_t} = \boldsymbol{z}'_{i_t}$, leading to $\boldsymbol{w}_t = \boldsymbol{w}'_t, \boldsymbol{v}_t = \boldsymbol{v}'_t$.

Then we focus on $\prod_{t=t_0}^{T}(\boldsymbol{I}_3 + \boldsymbol{R}_t^+)$. Notice that all elements in this product are nonnegative, thus in the remainder of this proof, given two nonnegative matrics $\boldsymbol{a}, \boldsymbol{b}$ with the same shape, we say $\boldsymbol{a} \leq \boldsymbol{b}$ if all elements in $\boldsymbol{a}$ are not larger than that in $\boldsymbol{b}$.

Denote

$$\boldsymbol{M} = LC_\eta\begin{bmatrix} 2 & (1-\beta) & 1/N^+ \\ \beta & 0 & 1/N^+ \\ 0 & 0 & 0 \end{bmatrix},$$

then by setting $\eta_t \leq C_\eta / t$ we have

$$
\begin{aligned}
&\prod_{t=t_0}^{T} (\boldsymbol{I}_3 + \boldsymbol{R}_t^+) \\
\leq &\prod_{t=t_0}^{T} \begin{bmatrix} 1+2L\eta_t & L\eta_t/N^+ & L\eta_t \\ L\beta & 1 & 1/N^+ \\ 0 & 0 & 1 \end{bmatrix} \\
= &\prod_{t=t_0}^{T} \begin{bmatrix} 1+2L\eta_t & L(1-\beta)\eta_t & L\eta_t \\ L\beta & N^+ & 1/N^+ \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
= &\begin{bmatrix} 1 & 0 & 0 \\ 0 & N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix} \prod_{t=t_0}^{T} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1+2L\eta_t & L(1-\beta)\eta_t & L\eta_t \\ L\beta & N^+ & 1/N^+ \\ 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
= &\begin{bmatrix} 1 & 0 & 0 \\ 0 & N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix} \prod_{t=t_0}^{T} \left(\boldsymbol{I}_3 + \begin{bmatrix} 2L\eta_t & L(1-\beta)\eta_t & L\eta_t \\ L\beta/N^+ & 0 & 1/(N^+)^2 \\ 0 & 0 & 0 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
\leq &\begin{bmatrix} 1 & 0 & 0 \\ 0 & N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix} \prod_{t=t_0}^{T} (\boldsymbol{I}_3 + \boldsymbol{M}/t) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix},
\end{aligned}
\tag{63}
$$

where the last inequation is due to $N^+ \geq T$. Notice that $\boldsymbol{M} \geq \boldsymbol{0}_{3\times 3}$, thus we have

$$
\begin{aligned}
\prod_{t=t_0}^{T} (\boldsymbol{I}_3 + \boldsymbol{M}/t) \leq &\prod_{t=t_0}^{T} \exp(\boldsymbol{M}/t) \\
= &\exp\left(\sum_{t=t_0}^{T} \boldsymbol{M}/t \right) \\
\leq &\exp\left(\boldsymbol{M} \log(T/(t_0-1)) \right) \\
= &\Lambda \; diag\left(\left(\frac{T}{t_0-1}\right)^{\lambda_1}, \left(\frac{T}{t_0-1}\right)^{\lambda_2}, \left(\frac{T}{t_0-1}\right)^{\lambda_3} \right) \Lambda^{-1},
\end{aligned}
\tag{64}
$$

where the last eqution is obtained by diagonalizing $\boldsymbol{M}$ into $\Lambda \; diag(\lambda_1, \lambda_2, \lambda_3)\Lambda^{-1}$, λ_i are the eigenvalues of $\boldsymbol{M}$ and the columns of Λ are the corresponding eigen vectors.

Combining Eq. (63) and Eq. (64), we have

$$
\begin{aligned}
&\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \prod_{t=t_0}^{T} (\boldsymbol{I}_3 + \boldsymbol{R}_t^+) \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^\top \\
\leq &\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix} \prod_{t=t_0}^{T} (\boldsymbol{I}_3 + \boldsymbol{M}/t) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/N^+ & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\
\leq &\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \prod_{t=t_0}^{T} (\boldsymbol{I}_3 + \boldsymbol{M}/t) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\
\leq &\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \Lambda \; diag\left(\left(\frac{T}{t_0-1}\right)^{\lambda_1}, \left(\frac{T}{t_0-1}\right)^{\lambda_2}, \left(\frac{T}{t_0-1}\right)^{\lambda_3} \right) \Lambda^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.
\end{aligned}
\tag{65}
$$

Specifically, we have $\lambda_{1,2} = LC_\eta(1 \pm \sqrt{1+\beta-\beta^2})$, $\lambda_3 = 0$, and

$$
\Lambda = \begin{bmatrix} 1 & 1 & \frac{1}{\beta N^+} \\ \frac{\beta(1-\sqrt{1+\beta-\beta^2})}{1-\beta} & \frac{\beta(1+\sqrt{1+\beta-\beta^2})}{1-\beta} & -\frac{2+\beta}{\beta(1-\beta)N^+} \\ 0 & 0 & -1 \end{bmatrix},
\tag{66}
$$

leading to

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \Lambda = \begin{bmatrix} 1 & 1 & \frac{1}{\beta N^+} \end{bmatrix}, \tag{67}$$

$$\Lambda^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^\top = \begin{bmatrix} \frac{2+2\beta+\beta\sqrt{1+\beta-\beta^2}}{2\beta^2\sqrt{1+\beta-\beta^2}N^+} \\ \frac{-2-2\beta+\beta\sqrt{1+\beta-\beta^2}}{2\beta^2\sqrt{1+\beta-\beta^2}N^+} \\ -1 \end{bmatrix}^\top, \tag{68}$$

$$\begin{aligned} &\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \Lambda \, diag\left(\left(\frac{T}{t_0-1}\right)^{\lambda_1}, \left(\frac{T}{t_0-1}\right)^{\lambda_2}, \left(\frac{T}{t_0-1}\right)^{\lambda_3}\right) \Lambda^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\ =& \frac{2+2\beta+\beta\sqrt{1+\beta-\beta^2}}{2\beta^2\sqrt{1+\beta-\beta^2}} \cdot \frac{T^{\lambda_1}}{(t_0-1)^{\lambda_1}N^+} \\ &+ \frac{-2-2\beta+\beta\sqrt{1+\beta-\beta^2}}{2\beta^2\sqrt{1+\beta-\beta^2}N^+} \cdot \frac{T^{\lambda_2}}{(t_0-1)^{\lambda_2}N^+} \\ &- \frac{T^{\lambda_2}}{\beta(t_0-1)^{\lambda_2}N^+} \\ \leq& \frac{1+2\beta}{\beta^2} \cdot \frac{T^{\lambda_1}}{(t_0-1)^{\lambda_1}N^+} \end{aligned} \tag{69}$$

Combining Eq. (62), Eq. (65), Eq. (69), we have

$$\frac{1}{N^+} \sum_{y_i=1} \mathbb{E}_{\mathcal{S},A} \left[\|\boldsymbol{w}_{T+1} - \boldsymbol{w}_{T+1}^{(i)}\|_2 \right] = \mathcal{O}\left((T/t_0)^{\lambda_1}/N^+ + \frac{t_0 n^+}{N^+} \right), \tag{70}$$

by choosing $t_0 \asymp T^{\frac{\lambda_1}{\lambda_1+1}} \cdot (n^+)^{-\frac{1}{\lambda_1+1}}$, we have

$$\frac{1}{N^+} \sum_{y_i=1} \mathbb{E}_{\mathcal{S},A} \left[\|\boldsymbol{w}_{T+1} - \boldsymbol{w}_{T+1}^{(i)}\|_2 \right] = \mathcal{O}\left(\frac{T^{\frac{\lambda_1}{\lambda_1+1}} \cdot (n^+)^{\frac{\lambda_1}{\lambda_1+1}}}{N^+} \right). \tag{71}$$

Similarly, for $y_i = -1$, let

$$\boldsymbol{M}' = LC_\eta \begin{bmatrix} 2 & (1-\beta) & 1/N^- \\ \beta & 0 & 1/N^- \\ 0 & 0 & 0 \end{bmatrix},$$

which shares the same eigenvalues with $\boldsymbol{M}$, thus we have

$$\frac{1}{N^-} \sum_{y_i=-1} \mathbb{E}_{\mathcal{S},A} \left[\|\boldsymbol{w}_{T+1} - \boldsymbol{w}_{T+1}^{(i)}\|_2 \right] = \mathcal{O}\left(\frac{T^{\frac{\lambda_1}{\lambda_1+1}} \cdot (n^-)^{\frac{\lambda_1}{\lambda_1+1}}}{N^-} \right). \tag{72}$$

□

C.3 Convergence Rate

Theorem 3. *Let Asmp. 1, 3, 4 hold. By setting $\eta_t = \frac{2t+1}{\mu(t+1)^2}$ and $\beta_t = \beta \asymp 1/n^+$, we have*

$$\mathbb{E}_A[F(\boldsymbol{w}_{T+1}) - F(\boldsymbol{w}^*)] = \mathcal{O}\left(n^+/T + 1/N^+\right). \tag{17}$$

Proof. First, according to the recurrence of $\boldsymbol{v}_t$, we have

$$\boldsymbol{v}_{t+1} = \sum_{\tau=1}^{t} (1-\beta)^{t-\tau} \beta \cdot \phi\left(h_{\boldsymbol{w}_\tau}(\boldsymbol{z}_{i_\tau})\right). \tag{73}$$

For sake of the presentation, denote

$$
\begin{aligned}
\boldsymbol{u}_{t+1} &= \sum_{\tau=1}^{t}(1-\beta)^{t-\tau}\beta \cdot h_{\boldsymbol{w}_\tau}(\mathcal{S}^+) \\
\boldsymbol{g}_t &= \mathbb{E}_{\mathcal{S}}[\nabla f\left(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}_{t+1}\right)] \\
\delta_1^t &= \nabla F(\boldsymbol{w}_t) - \mathbb{E}_{\mathcal{S}}\left[\nabla f\left(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, h_{\boldsymbol{w}_t}(\mathcal{S}^+)\right)\right] \\
\delta_2^t &= \mathbb{E}_{\mathcal{S}}\left[\nabla f\left(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, h_{\boldsymbol{w}_t}(\mathcal{S}^+)\right) - \nabla f\left(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{u}_{t+1}\right)\right] \\
\delta_3^t &= \mathbb{E}_{\mathcal{S}}\left[\nabla f\left(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{u}_{t+1}\right) - \nabla f\left(\boldsymbol{w}_t; \boldsymbol{z}_{i_t}, \boldsymbol{v}_{t+1}\right)\right].
\end{aligned}
\tag{74}
$$

According to the smoothness of F, we have

$$
\begin{aligned}
F(\boldsymbol{w}_{t+1}) \leq & F(\boldsymbol{w}_t) + \langle \nabla F(\boldsymbol{w}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle + \frac{L_w}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_2^2 \\
= & F(\boldsymbol{w}_t) - \eta_t \langle \nabla F(\boldsymbol{w}_t), \boldsymbol{g}_t \rangle + \frac{L_w \eta_t^2}{2}\|\boldsymbol{g}_t\|_2^2 \\
= & F(\boldsymbol{w}_t) - \eta_t \|\nabla F(\boldsymbol{w}_t)\|_2^2 + \eta_t \langle \nabla F(\boldsymbol{w}_t), \delta_1^t + \delta_2^t + \delta_3^t \rangle + \frac{L_w \eta_t^2}{2}\|\boldsymbol{g}_t\|_2^2.
\end{aligned}
\tag{75}
$$

By taking expectation of A on both sides, we have

$$
\begin{aligned}
\mathbb{E}_A[F(\boldsymbol{w}_{t+1})] \leq & \mathbb{E}_A[F(\boldsymbol{w}_t)] - \eta_t \mathbb{E}_A[\|\nabla F(\boldsymbol{w}_t)\|_2^2] + \frac{L_w \eta_t^2}{2}\mathbb{E}_A\left[\|\boldsymbol{g}_t\|_2^2\right] \\
& + \eta_t \mathbb{E}_A\left[\langle \nabla F(\boldsymbol{w}_t), \delta_2^t + \delta_3^t \rangle\right] \\
\leq & \mathbb{E}_A[F(\boldsymbol{w}_t)] - \eta_t \mathbb{E}_A[\|\nabla F(\boldsymbol{w}_t)\|_2^2] + \frac{L_w \eta_t^2}{2}\mathbb{E}_A\left[\|\boldsymbol{g}_t\|_2^2\right] \\
& + \frac{\eta_t}{2}\mathbb{E}_A\left[\|\delta_2^t + \delta_3^t\|_2^2\right] + \frac{\eta_t}{2}\mathbb{E}_A[\|\nabla F(\boldsymbol{w}_t)\|_2^2] \\
\leq & \mathbb{E}_A[F(\boldsymbol{w}_t)] - \frac{\eta_t}{2}\mathbb{E}_A[\|\nabla F(\boldsymbol{w}_t)\|_2^2] + \frac{L_w G^2 \eta_t^2}{2} \\
& + \eta_t \mathbb{E}_A\left[\|\delta_2^t\|_2^2 + \|\delta_3^t\|_2^2\right].
\end{aligned}
\tag{76}
$$

By using the PL condition and subtracting $F(\boldsymbol{w}^*)$ from both sides, we get

$$
\begin{aligned}
\mathbb{E}_A[F(\boldsymbol{w}_{t+1}) - F(\boldsymbol{w}^*)] \leq & \left(1 - \frac{\mu \cdot \eta_t}{2}\right)\mathbb{E}_A[F(\boldsymbol{w}_t) - F(\boldsymbol{w}^*)] + \frac{L_w G^2 \eta_t^2}{2} \\
& + \eta_t \mathbb{E}_A\left[\|\delta_2^t\|_2^2 + \|\delta_3^t\|_2^2\right].
\end{aligned}
\tag{77}
$$

If we set the learning rate $\eta_t = \frac{2(2t+1)}{\mu(t+1)^2}$, we have $1 - \mu \cdot \eta_t/2 = t^2/(t+1)^2$, leading to the follwing result:

$$
\begin{aligned}
& (t+1)^2 \cdot \mathbb{E}_A[F(\boldsymbol{w}_{t+1}) - F(\boldsymbol{w}^*)] \\
\leq & T^2 \mathbb{E}_A[F(\boldsymbol{w}_t) - F(\boldsymbol{w}^*)] + \frac{2(2t+1)^2 L_w G^2}{\mu^2 (t+1)^2} + \frac{2(2t+1)}{\mu}\mathbb{E}_A\left[\|\delta_2^t\|_2^2 + \|\delta_3^t\|_2^2\right] \\
\leq & t^2 \mathbb{E}_A[F(\boldsymbol{w}_t) - F(\boldsymbol{w}^*)] + \frac{8 L_w G^2}{\mu^2} + \frac{2(2t+1)}{\mu}\mathbb{E}_A\left[\|\delta_2^t\|_2^2 + \|\delta_3^t\|_2^2\right].
\end{aligned}
\tag{78}
$$

Applying the above result recursively, we get

$$
\begin{aligned}
& (T+1)^2 \cdot \mathbb{E}_A[F(\boldsymbol{w}_{t+1}) - F(\boldsymbol{w}^*)] \\
\leq & \mathbb{E}_A[F(\boldsymbol{w}_1) - F(\boldsymbol{w}^*)] + \frac{8 L_w G^2 T}{\mu^2} + \sum_{t=2}^{T}\frac{2(2t+1)}{\mu}\mathbb{E}_A\left[\|\delta_2^t\|_2^2 + \|\delta_3^t\|_2^2\right] \\
\leq & \mathbb{E}_A[F(\boldsymbol{w}_1) - F(\boldsymbol{w}^*)] + \frac{8 L_w G^2 T}{\mu^2} + \sum_{t=2}^{T}\frac{4(t+1)}{\mu}\mathbb{E}_A\left[\|\delta_2^t\|_2^2 + \|\delta_3^t\|_2^2\right].
\end{aligned}
\tag{79}
$$

Moreover, according to Lem. 4, we have

$$\begin{aligned}\sum_{t=2}^{T}(t+1)\mathbb{E}_A\left[\|\delta_2^t\|_2^2\right] \leq& \left(\frac{L_r L_2 C_w G}{\beta N^+}\right)^2 \sum_{t=2}^{T}\frac{t+1}{t-1}\sum_{\tau=1}^{t-1}\frac{4(2\tau+1)^2}{\mu^2(\tau+1)^4}\\ \leq& 48\left(\frac{L_r L_2 C_w G}{\mu\beta N^+}\right)^2 \sum_{t=2}^{T}\sum_{\tau=1}^{t-1}\frac{1}{(\tau+1)^2}\\ \leq& 24\left(\frac{L_r L_2 C_w G}{\mu\beta N^+}\right)^2 (T-1),\end{aligned} \tag{80}$$

and

$$\sum_{t=2}^{T}(t+1)\mathbb{E}_A\left[\|\delta_3^t\|_2^2\right] \leq \left(\frac{L_r^2}{N^+}+\frac{L_2^2 L_r^2\kappa^2}{(N^+)^2}\right)\cdot\frac{(T+4)(T-1)}{2}. \tag{81}$$

From Eq. (79), Eq. (80), Eq. (81), we have

$$\begin{aligned}&\mathbb{E}_A[F(\boldsymbol{w}_{t+1})-F(\boldsymbol{w}^*)]\\ \leq&\frac{\mathbb{E}_A[F(\boldsymbol{w}_1)-F(\boldsymbol{w}^*)]}{(T+1)^2}+\frac{8L_w G^2\mu+96(L_r L_2 C_w G)^2/\beta^2}{\mu^3(T+1)}+\frac{4L_r^2+4L_2^2L_r^2\kappa^2}{\mu N^+}.\end{aligned} \tag{82}$$

□

Lemma 4. *Denote $r(s)=\sigma\left((1-\pi)/\pi\cdot\hat{\mathbb{E}}_{\boldsymbol{x}\sim\boldsymbol{z}^-}[\ell_1(h_{\boldsymbol{w}}(\boldsymbol{x}^+)-h_{\boldsymbol{w}}(\boldsymbol{x}))]/s\right)$. Assume $\nabla r(s)$ is a L_r-Lipschitz smooth function with regard to s, and ℓ_2 is L_2-Lipschitz smooth, then the following facts hold:*

(a) $\|\nabla f(\boldsymbol{w};\boldsymbol{z},\boldsymbol{v}_1)-\nabla f(\boldsymbol{w};\boldsymbol{z},\boldsymbol{v}_2)\|_2 \leq \frac{L_r L_2}{N^+}\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2$.

(b) $\|\delta_2^t\|_2^2 \leq \left(\frac{L_r L_2 C_w G}{\beta N^+}\right)^2\cdot\sum_{\tau=1}^{t-1}\eta_\tau^2/(t-1)$.

(c) Assume $\mathbb{E}_{\mathcal{S}}\left[\|h_{\boldsymbol{w}}(\mathcal{S})-\phi(h_{\boldsymbol{w}}(\boldsymbol{z}))\|_2\right]$ is upper bounded by κ, then we have $\mathbb{E}_A\left[\|\delta_3^t\|_2^2\right] \leq \frac{L_r^2}{N^+}+\frac{L_2^2L_r^2\kappa^2}{(N^+)^2}$.

Proof. First, we have $\nabla f(\boldsymbol{w};\boldsymbol{z},\boldsymbol{v}) = \hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[r\left(\hat{\mathbb{E}}_{v\sim\boldsymbol{v}}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v)]\right)\right]$, according to the Lipschitz smoothness of ∇r, we have

$$\begin{aligned}&\|\nabla f(\boldsymbol{w};\boldsymbol{z},\boldsymbol{v}_1)-\nabla f(\boldsymbol{w};\boldsymbol{z},\boldsymbol{v}_1)\|_2\\ =&\left\|\hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\nabla r\left(\hat{\mathbb{E}}_{v\sim\boldsymbol{v}_1}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v)]\right)-\nabla r\left(\hat{\mathbb{E}}_{v'\sim\boldsymbol{v}_2}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v')]\right)\right]\right\|_2\\ \leq&\hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\left\|\nabla r\left(\hat{\mathbb{E}}_{v\sim\boldsymbol{v}_1}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v)]\right)-\nabla r\left(\hat{\mathbb{E}}_{v'\sim\boldsymbol{v}_2}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v')]\right)\right\|_2\right]\\ \leq& L_r\cdot\hat{\mathbb{E}}_{\boldsymbol{z},c\sim h_{\boldsymbol{w}}(\boldsymbol{z}^+)}\left[\left|\hat{\mathbb{E}}_{v\sim\boldsymbol{v}_1}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v)]-\hat{\mathbb{E}}_{v'\sim\boldsymbol{v}_2}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v')]\right|\right].\end{aligned} \tag{83}$$

Denote the i-th element of $\boldsymbol{v}$ as $\boldsymbol{v}^{(i)}$, for any $\boldsymbol{x}$ we have

$$\begin{aligned}
&\left|\hat{\mathbb{E}}_{v\sim\boldsymbol{v}_1}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v)]-\hat{\mathbb{E}}_{v'\sim\boldsymbol{v}_2}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v')]\right| \\
=&\left|\frac{1}{N^+}\sum_{i=1}^{N^+}\left(\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-\boldsymbol{v}_1^{(i)})-\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-\boldsymbol{v}_2^{(i)})\right)\right| \\
=&\frac{1}{N^+}\sum_{i=1}^{N^+}\left|\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-\boldsymbol{v}_1^{(i)})-\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-\boldsymbol{v}_2^{(i)})\right| \\
\leq&\frac{L_2}{N^+}\cdot\sum_{i=1}^{N^+}\left|\boldsymbol{v}_1^{(i)}-\boldsymbol{v}_2^{(i)}\right| \\
=&\frac{L_2}{N^+}\cdot\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_1 \\
\leq&\frac{L_2}{N^+}\cdot\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2,
\end{aligned} \tag{84}$$

Eq. (83) and Eq. (84) leads to conclusion *(a)* imediately.

We then prove conclusion *(b)*. According to the above derivation, we have

$$\|\delta_2^t\|_2^2\leq\left(\frac{L_rL_c}{N^+}\right)^2\cdot\left\|h_{\boldsymbol{w}_t}(\mathcal{S}^+)-\boldsymbol{u}_{t+1}\right\|_2^2, \tag{85}$$

where the right side could be upper bounded by the following recursion

$$\begin{aligned}
&\left\|h_{\boldsymbol{w}_t}(\mathcal{S}^+)-\boldsymbol{u}_{t+1}\right\|_2^2 \\
=&(1-\beta)^2\left\|h_{\boldsymbol{w}_t}(\mathcal{S}^+)-\boldsymbol{u}_t\right\|_2^2 \\
\leq&(1-\beta)^2\left\|h_{\boldsymbol{w}_t}(\mathcal{S}^+)-h_{\boldsymbol{w}_{t-1}}(\mathcal{S}^+)+h_{\boldsymbol{w}_{t-1}}(\mathcal{S}^+)-\boldsymbol{u}_t\right\|_2^2 \\
\overset{(*)}{\leq}&(1-\beta)^2(1+\beta)\left\|h_{\boldsymbol{w}_{t-1}}(\mathcal{S}^+)-\boldsymbol{u}_t\right\|_2^2 \\
&\qquad+(1-\beta)^2(1+\beta^{-1})\left\|h_{\boldsymbol{w}_t}(\mathcal{S}^+)-h_{\boldsymbol{w}_{t-1}}(\mathcal{S}^+)\right\|_2^2 \\
\leq&(1-\beta)\left\|h_{\boldsymbol{w}_{t-1}}(\mathcal{S}^+)-\boldsymbol{u}_t\right\|_2^2+(1-\beta)\beta^{-1}\left\|h_{\boldsymbol{w}_t}(\mathcal{S}^+)-h_{\boldsymbol{w}_{t-1}}(\mathcal{S}^+)\right\|_2^2 \\
\leq&(1-\beta)\left\|h_{\boldsymbol{w}_{t-1}}(\mathcal{S}^+)-\boldsymbol{u}_t\right\|_2^2+(1-\beta)\beta^{-1}C_w^2\left\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\right\|_2^2 \\
=&(1-\beta)\left\|h_{\boldsymbol{w}_{t-1}}(\mathcal{S}^+)-\boldsymbol{u}_t\right\|_2^2+(1-\beta)\beta^{-1}C_w^2\eta_{t-1}^2\left\|\boldsymbol{g}_{t-1}\right\|_2^2.
\end{aligned} \tag{86}$$

Step (*) is due to the Cauchy-Schwarz inequality. By iterating the above recursion, we get

$$\begin{aligned}
\left\|h_{\boldsymbol{w}_t}(\mathcal{S}^+)-\boldsymbol{u}_{t+1}\right\|_2^2\leq&C_w^2\sum_{\tau=1}^{t-1}(1-\beta)^{t-\tau}\beta^{-1}\eta_\tau^2\|\boldsymbol{g}_\tau\|_2^2 \\
\leq&C_w^2G^2\sum_{\tau=1}^{t-1}(1-\beta)^{t-\tau}\beta^{-1}\eta_\tau^2 \\
\leq&\frac{C_w^2G^2}{t-1}\sum_{\tau=1}^{t-1}(1-\beta)^{t-\tau}\beta^{-1}\cdot\sum_{\tau=1}^{t-1}\eta_\tau^2 \\
\leq&\frac{C_w^2G^2}{\beta^2(t-1)}\cdot\sum_{\tau=1}^{t-1}\eta_\tau^2
\end{aligned} \tag{87}$$

The proof of *(b)* is complete due to Eq. (85) and Eq. (87).

Now we turn to conclusion *(c)*. First of all, denote

$$\Delta(\mathcal{S})=\hat{\mathbb{E}}_{v\sim h_{\boldsymbol{w}}(\mathcal{S})}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v)]-\hat{\mathbb{E}}_{v'\sim\phi(h_{\boldsymbol{w}}(\boldsymbol{z}))}[\ell_2(h_{\boldsymbol{w}}(\boldsymbol{x})-v')], \tag{88}$$

where $\mathcal{S}$ is a dataset with size N and $\boldsymbol{z}$ is a randomly sampled subset of $\mathcal{S}$, we will show that for any $\boldsymbol{w} \in \Omega$, $\mathbb{E}_{\mathcal{S}}\left[\|\Delta(\mathcal{S})\|_2^2\right] = \mathcal{O}\left(1/N^+\right)$. Consider a dataset $\mathcal{S}'$ with at most one example $\boldsymbol{x}_0$ differs from $\mathcal{S}$, $\Delta(\mathcal{S})$ satisifies the bounded differences condition:

$$\begin{aligned}
&\mathbb{E}_A\left[\sup_{\mathcal{S},\mathcal{S}'}|\Delta(\mathcal{S})-\Delta(\mathcal{S}')|\right] \\
=&\sup_{\mathcal{S},\mathcal{S}'}\left[|\Delta(\mathcal{S})-\Delta(\mathcal{S}')|\,\middle|\boldsymbol{x}_0\in\boldsymbol{z}\right]\cdot\mathbb{P}\left[\boldsymbol{x}_0\in\boldsymbol{z}\right] \\
&+\sup_{\mathcal{S},\mathcal{S}'}\left[|\Delta(\mathcal{S})-\Delta(\mathcal{S}')|\,\middle|\boldsymbol{x}_0\notin\boldsymbol{z}\right]\cdot\mathbb{P}\left[\boldsymbol{x}_0\notin\boldsymbol{z}\right] \\
\leq&\frac{1}{n^+}\cdot\frac{n^+}{N^+}+\frac{1}{N^+}\cdot\frac{N^+-n^+}{N^+} \\
\leq&\frac{2}{N^+}.
\end{aligned} \tag{89}$$

According to Lem. 5, we have

$$\mathbb{E}_A\left[Var_{\mathcal{S}}(\Delta(\mathcal{S}))\right] \leq \frac{1}{4}\cdot N^+\cdot\left(\frac{2}{N^+}\right)^2 = \frac{1}{N^+}. \tag{90}$$

Furthermore, similar to Eq. (84), we have

$$\mathbb{E}_{\mathcal{S}}[\Delta(\mathcal{S})] \leq \frac{L_2}{N^+}\mathbb{E}_{\mathcal{S}}\left[\|h_{\boldsymbol{w}}(\mathcal{S})-\phi(h_{\boldsymbol{w}}(\boldsymbol{z}))\|_2\right] \leq \frac{L_2\kappa}{N^+}, \tag{91}$$

leading to

$$\begin{aligned}
\mathbb{E}_A\left[\|\delta_3\|_2^2\right] \leq& L_r\mathbb{E}_{\mathcal{S}}\left[|\Delta(\mathcal{S})|_2^2\right] \\
\leq& L_r^2\mathbb{E}_A\left[Var_{\mathcal{S}}(\Delta(\mathcal{S}))\right]+L_r^2\mathbb{E}_A\left[\mathbb{E}_{\mathcal{S}}^2[\Delta(\mathcal{S})]\right] \\
\leq& \frac{L_r^2}{N^+}+\frac{L_2^2L_r^2\kappa^2}{(N^+)^2}.
\end{aligned} \tag{92}$$

□

Lemma 5 (Efron-Stein Inequality [6]). *Let $X_1,\cdots,X_n$ be independent random variables and let $Z=f(X_1,\cdots,X_n)$, if f has the bounded differences property with constants $c_1,\cdots,c_n$, then*

$$Var(Z) \leq \frac{1}{4}\sum_{k=1}^{n}c_k^2. \tag{93}$$

D Experiments

D.1 Implementation Details

Competitors. To validate the advantages of the proposed method over the state-of-the art methods in image retrieval, we compare two types of competitors: **1) Pairwise Losses**, including *Contrastive Loss* [27], *Triplet Loss* [32], *Multi-Similarity (MS) Loss* [71], *Cross-Batch Memory (XBM)* [72]. These methods construct loss functions with image pairs or triplets. **2) Ranking-Based Losses**, including *SmoothAP* [9], *FastAP* [12], *DIR* [58], *BlackBox* [33], and *Area Under the ROC Curve Loss (AUROC)* [76]. These methods directly optimize the ranking-based metrics such as AUPRC or AUROC. We reimplement all the competitors on the same codebase to ensure that the main experimental setting of competitors is the same as ours, including model structure, data preprocessing and augmentation, learning rate schedule, testing pipeline, etc. The unique hyperparameters of competitors follows the optimal settings of the original paper. Moreover, the optimizer used are slightly different: following previous work, Adam is used to train the competitors, while ours is trained with SGD to ensure the consistence with our theoretical analysis. We also provide the results of ours trained with Adam in Tab. 2, which shows no significant difference.

Environments. The proposed method and all competitors are implemented with Pytorch 1.8.2 [51]. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

D.2 More Results

Quantitative results. We provide full results with more k for the Recall@k metric in Tab. 3, 5, 7. In addition, since the mainstream work on image retrieval does not split a validation but instead trains models with all data except the test data (called *trainval* split here), we further provide evaluation results of some methods trained on the *trainval* split in Tab. 4, 6, 8.

Table 3: Quantitative results on SOP. All methods are trained on the *train split*.

Methods	mAUPRC	R@1	R@10	R@100	R@1000
Contrastive loss [27]	57.73	77.60	89.31	95.54	98.65
Triplet loss [32]	58.07	78.34	90.50	96.20	98.88
MS loss [71]	60.10	79.64	90.38	95.93	98.72
XBM [72]	61.29	80.66	91.08	96.04	98.55
SmoothAP [9]	**61.65**	**81.13**	**92.02**	**96.69**	**98.91**
DIR [58]	60.74	80.52	91.35	96.46	98.80
FastAP [12]	57.10	77.30	89.61	95.74	98.72
AUROC [25]	55.80	77.32	89.64	95.77	98.76
BlackBox [52]	59.74	79.48	90.74	96.16	98.74
Ours	**62.75**	**81.91**	**92.50**	**96.97**	**98.96**

Table 4: Quantitative results on SOP. All methods are trained on the *trainval split*.

Methods	mAUPRC	R@1	R@10	R@100	R@1000
Triplet loss [32]	58.34	78.61	90.73	96.33	98.90
SmoothAP [9]	**61.96**	**81.18**	**92.10**	**96.70**	**98.97**
DIR [58]	61.29	80.74	91.59	96.58	98.89
FastAP [12]	57.39	77.56	89.79	95.88	98.69
AUROC [25]	55.97	77.42	89.65	95.83	98.72
Ours	**63.27**	**82.15**	**92.69**	**97.11**	**99.02**

Table 5: Quantitative results on iNaturalist. All methods are trained on the *train split*.

Methods	mAUPRC	R@1	R@4	R@16	R@32
Contrastive loss [27]	27.99	54.19	71.12	82.77	86.94
Triplet loss [32]	30.59	60.53	77.62	88.20	91.72
MS loss [71]	30.28	63.39	78.50	88.07	91.40
XBM [72]	27.46	59.12	75.18	85.93	89.72
SmoothAP [9]	**33.92**	**66.13**	**80.93**	**89.71**	**92.67**
DIR [58]	33.51	64.86	79.79	89.07	92.20
FastAP [12]	31.02	56.64	73.57	84.65	88.49
AUROC [25]	27.24	60.88	77.76	88.39	91.85
BlackBox [52]	29.28	56.88	74.10	85.41	89.42
Ours	**36.16**	**68.22**	**82.86**	**91.02**	**93.71**

Qualitative results. We show more mean PR curves in Fig. 6 and more convergence results in Fig. 7. These qualitative results are consistent with our main conclusions.

Table 6: Quantitative results on iNaturalist. All methods are trained on the *trainval split*.

Methods	mAUPRC	R@1	R@4	R@16	R@32
Triplet loss [32]	31.49	61.10	77.97	88.47	91.95
SmoothAP [9]	**34.89**	**66.79**	**81.55**	**90.13**	**92.95**
DIR [58]	34.55	65.65	80.63	89.51	92.52
FastAP [12]	32.14	57.84	74.59	85.27	88.99
AUROC [25]	26.08	58.70	76.13	87.33	91.05
Ours	**37.31**	**69.23**	**83.41**	**91.29**	**93.81**

Table 7: Quantitative results on VehicleID. All methods are trained on the *train split*.

Methods	Small			Medium			Large		
	mAUPRC	R@1	R@5	mAUPRC	R@1	R@5	mAUPRC	R@1	R@5
Contrastive loss [27]	77.74	92.08	92.08	72.33	90.15	95.40	67.26	87.46	94.60
Triplet loss [32]	80.27	94.18	**97.54**	75.58	92.10	96.20	70.99	90.09	95.54
MS loss [71]	78.83	93.05	97.00	74.30	91.46	96.06	69.15	88.82	95.06
XBM [72]	77.45	**94.83**	96.94	73.77	**93.63**	96.01	71.24	**92.78**	95.83
SmoothAP [9]	80.39	94.36	97.44	76.28	92.98	96.42	72.28	91.31	96.05
DIR [58]	**81.17**	94.64	97.43	**76.95**	92.74	**96.45**	**72.72**	91.38	**96.10**
FastAP [12]	80.67	93.67	97.10	75.73	91.58	96.29	70.82	89.42	95.38
AUROC [25]	68.27	88.31	95.26	63.41	85.57	93.42	58.12	81.73	91.92
BlackBox [52]	80.00	93.72	97.15	75.46	91.84	96.26	70.92	90.14	95.52
Ours	**83.11**	**95.41**	**97.69**	**79.00**	**93.67**	**96.68**	**74.95**	**92.50**	**96.44**

Table 8: Quantitative results on VehicleID. All methods are trained on the *trainval split*.

Methods	Small			Medium			Large		
	mAUPRC	R@1	R@5	mAUPRC	R@1	R@5	mAUPRC	R@1	R@5
Triplet loss [32]	80.73	94.01	97.41	75.48	92.14	96.28	70.75	89.65	95.53
SmoothAP [9]	80.74	94.39	97.34	76.76	92.90	96.36	72.65	91.49	96.06
DIR [58]	**81.59**	**94.69**	**97.43**	**77.36**	**93.15**	**96.46**	**73.20**	**91.74**	**96.14**
FastAP [12]	81.04	93.35	97.35	75.98	91.58	96.01	70.95	89.39	95.48
AUROC [25]	68.39	88.34	95.43	63.55	85.82	93.42	58.25	81.81	91.92
Ours	**83.70**	**95.41**	**97.72**	**79.84**	**93.80**	**96.77**	**75.56**	**92.83**	**96.48**

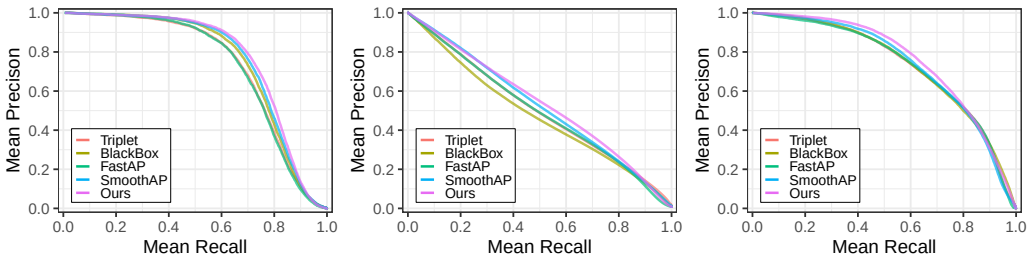


Figure 6: Mean PR curves of different methods on validation sets of SOP (left), iNaturalist (middle), and VehicleID (right).

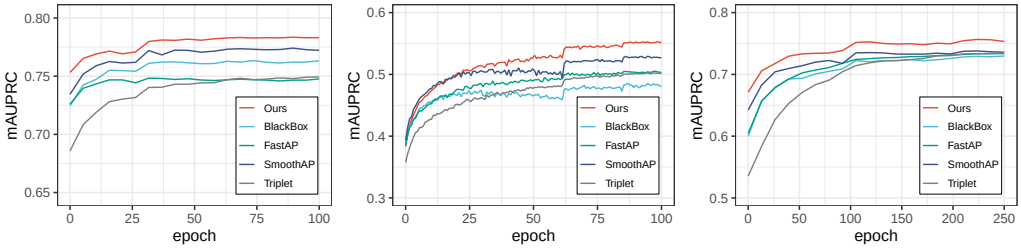


Figure 7: Convergence results of different methods on validation sets of SOP (left), iNaturalist (middle), and VehicleID (right).