Risk Bounds for Low Cost Bipartite Ranking

San Gultekin and John Paisley

November 24, 2019

Abstract

Bipartite ranking is an important supervised learning problem; however, unlike regression or classification, it has a quadratic dependence on the number of samples. To circumvent the prohibitive sample cost, many recent work focus on stochastic gradient-based methods. In this paper we consider an alternative approach, which leverages the structure of the widely-adopted pairwise squared loss, to obtain a stochastic and low cost algorithm that does not require stochastic gradients or learning rates. Using a novel uniform risk bound—based on matrix and vector concentration inequalities—we show that the sample size required for competitive performance against the all-pairs batch algorithm does not have a quadratic dependence. Generalization bounds for both the batch and low cost stochastic algorithms are presented. Experimental results show significant speed gain against the batch algorithm, as well as competitive performance against state-of-the-art bipartite ranking algorithms on real datasets.

1 Introduction

Binary classification is among the most widely studied machine learning problems, with many applications. Given a binary labeled dataset, the aim is to learn a mapping from the features to the labels. The performance of a learning algorithm is typically gauged in terms of classification error. However, in situations such as cost-sensitive learning [1] and imbalanced learning [2, 3] this choice may not be appropriate. For example, in online advertising [1] one is typically concerned with separating the interesting ads from the rest. This problem is also known as bipartite ranking, where the aim is to rank the "positive" inputs higher than the "negative" ones.

1.1 Problem Setup

Let $\mathcal{X}$ be a D-dimensional input domain and $\mathcal{Y}$ the label domain. The bipartite ranking problem is defined for binary labeled samples, $\mathcal{Y} = \{0, 1\}$. These sample pairs are generated i.i.d. from

an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. A ranking function $f : \mathcal{X} \to \mathbb{R}$ is a mapping from the inputs to a scalar-valued score. In this paper we are interested in linear ranking functions of the form $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ with parameter vector $\boldsymbol{w}$. The goal in bipartite ranking is to assign higher scores to the inputs with label 1. Let $\boldsymbol{x}^1$ and $\boldsymbol{x}^0$ denote two samples with corresponding labels of 1 and 0. In many ranking settings we are given a set of features that encodes the preference of one sample over another [1], which can be represented by $\Psi(\boldsymbol{x}^1, \boldsymbol{x}^0)$. It is also possible that the individual features $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ are not defined explicitly; for example when a single user is shown two ads and prefers one over another. Given this, a widely used performance metric is the Wilcoxon-Mann-Whitney statistic

$$\boldsymbol{\ell}_{\mathrm{WMW}}(\boldsymbol{w}, \Psi(\boldsymbol{x}^1, \boldsymbol{x}^0)) = \mathbb{I}\left\{\boldsymbol{w}^\top \Psi(\boldsymbol{x}^1, \boldsymbol{x}^0) > 0\right\} + \frac{1}{2}\mathbb{I}\left\{\boldsymbol{w}^\top \Psi(\boldsymbol{x}^1, \boldsymbol{x}^0) = 0\right\} \ . \tag{1}$$

Based on this we define the risks

$$R^{\mathrm{AUC}}(\boldsymbol{w}) = \mathbb{E}_{\substack{x^1 \sim D^1 \\ x^0 \sim D^0}}\left[\boldsymbol{\ell}_{\mathrm{WMW}}(\boldsymbol{w}, \Psi(\boldsymbol{x}^1, \boldsymbol{x}^0))\right], \quad R^{\mathrm{AUC}}_{\mathcal{N}}(\boldsymbol{w}) = \frac{1}{N_1 N_0}\sum_{i=1}^{N_1}\sum_{j=1}^{N_0} \boldsymbol{\ell}_{\mathrm{WMW}}(\boldsymbol{w}, \Psi(\boldsymbol{x}_i^1, \boldsymbol{x}_j^0)), \tag{2}$$

where $R^{\mathrm{AUC}}(\boldsymbol{w})$ corresponds to the actual AUC risk and is obtained by taking expectation over the class-conditional distributions $D^i = D(\boldsymbol{x}|y = i)$. Let the sample set $\mathcal{N} = \{\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_{N_0}^0, \boldsymbol{x}_1^1, \ldots, \boldsymbol{x}_{N_1}^1\}$, containing N_0 and N_1 samples with labels 0 and 1, respectively, and define $N := N_0 + N_1$. This gives the empirical risk $R^{\mathrm{AUC}}_{\mathcal{N}}(\boldsymbol{w})$ to be minimized in practice. The main drawback of this approach is that the objective function is now a sum of indicator functions, an NP-hard problem. A widely-used approach to handle this problem is to replace the intrinsic loss function $\boldsymbol{\ell}_{\mathrm{WMW}}$ with a convex one, often written as $\boldsymbol{\ell}_\phi$ [4, 5].

In this paper we are interested in the pairwise squared loss as it is a consistent estimator of AUC [6], and widely preferred by recent work [7, 8, 9]:

$$\boldsymbol{\ell}_\phi(\boldsymbol{w}, (\boldsymbol{x}^1, \boldsymbol{x}^0)) = \frac{1}{2}\left[1 - \boldsymbol{w}^\top \Psi(\boldsymbol{x}_i^1, \boldsymbol{x}_j^0)\right]^2, \tag{3}$$

which produces its own corresponding actual and empirical ϕ-risks that parallels Eq. (2). Let $\boldsymbol{\Sigma}_N = 1/(N_1 N_0)\sum_{i=1}^{N_1}\sum_{j=1}^{N_0} \Psi(\boldsymbol{x}_i^1, \boldsymbol{x}_j^0)\Psi(\boldsymbol{x}_i^1, \boldsymbol{x}_j^0)^\top$ and $\boldsymbol{\mu}_N = 1/(N_1 N_0)\sum_{i=1}^{N_1}\sum_{j=1}^{N_0} \Psi(\boldsymbol{x}_i^1, \boldsymbol{x}_j^0)$. The empirical risk for this problem can then also written as the optimization of $R^\phi_{\mathcal{N}}(\boldsymbol{w}) = (1/2)\boldsymbol{w}^\top \boldsymbol{\Sigma}_N \boldsymbol{w} - \boldsymbol{\mu}_N^\top \boldsymbol{w}$. For what follows we will drop the ϕ symbol and refer to $\boldsymbol{\ell}_\phi$, R^ϕ, and $R^\phi_{\mathcal{N}}$ simply as $\boldsymbol{\ell}$, R, and $R_{\mathcal{N}}$.

The focus of this paper is on two algorithms: The first one is a Batch Bipartite Ranking (BBR) algorithm (Algorithm 1), which aims at minimizing $R_{\mathcal{N}}$ based on all pairs available in $\mathcal{N}$. We provide a new theoretical analysis of BBR in Section 2. Due to the quadratic growth of sample size, $O(N_1 N_0)$, the sample cost of BBR quickly becomes prohibitive. Therefore we also propose a new Low Cost Bipartite Ranking (LCBR) algorithm (Algorithm 2), which, given the same sample set $\mathcal{N}$, subsamples S pairs uniformly at random with replacement. The main goal of this paper is to analyze the subsample size S required for LCBR to be competitive with BBR. As we show in Section 3, S does not have such quadratic dependence. Section 4 discusses related work. We show experiments in Section 5 and conclude in Section 6.

Algorithm 1 BBR

Input: Sample set $\mathcal{N}$
Regularization parameter (W_*)
Output: Linear ranker's weight $\boldsymbol{w}_N$
1. Initialize $\boldsymbol{\mu}_N \leftarrow \mathbf{0}$, $\boldsymbol{\Sigma}_N \leftarrow \mathbf{0}$
2. `//Accummulation`
3. **for** $i = 1, \ldots, N_1$
4. **for** $j = 1, \ldots, N_0$
5. $\boldsymbol{\mu}_N \leftarrow \boldsymbol{\mu}_N + \frac{1}{N_1 N_0}(\boldsymbol{x}_i^1 - \boldsymbol{x}_j^0)$
6. $\boldsymbol{\Sigma}_N \leftarrow \boldsymbol{\Sigma}_N + \frac{1}{N_1 N_0}(\boldsymbol{x}_i^1 - \boldsymbol{x}_j^0)(\boldsymbol{x}_i^1 - \boldsymbol{x}_j^0)^\top$
7. **end for**
8. **end for**
9. `//Empirical Risk Minimization`
10. $\boldsymbol{w}_N \leftarrow \underset{\boldsymbol{w} \in B_2(W_*)}{\arg\min} \ \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{\Sigma}_N \boldsymbol{w} - \boldsymbol{\mu}_N^\top \boldsymbol{w}$

Algorithm 2 LCBR

Input: Sample set $\mathcal{N}$, subsample size (S)
Regularization parameter (W_*)
Output: Linear ranker's weight $\boldsymbol{w}_S$
1. Initialize $\boldsymbol{\mu}_S \leftarrow \mathbf{0}$, $\boldsymbol{\Sigma}_S \leftarrow \mathbf{0}$
2. `//Accummulation`
3. **for** $s = 1, \ldots, S$
4. Sample (i_s, j_s) uniformly with replacement
5. $\boldsymbol{\mu}_S \leftarrow \boldsymbol{\mu}_S + \frac{1}{S}(\boldsymbol{x}_{i_s}^1 - \boldsymbol{x}_{j_s}^0)$
6. $\boldsymbol{\Sigma}_S \leftarrow \boldsymbol{\Sigma}_S + \frac{1}{S}(\boldsymbol{x}_{i_s}^1 - \boldsymbol{x}_{j_s}^0)(\boldsymbol{x}_{i_s}^1 - \boldsymbol{x}_{j_s}^0)^\top$
7. **end for**
8.
9. `//Empirical Risk Minimization`
10. $\boldsymbol{w}_S \leftarrow \underset{\boldsymbol{w} \in B_2(W_*)}{\arg\min} \ \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{\Sigma}_S \boldsymbol{w} - \boldsymbol{\mu}_S^\top \boldsymbol{w}$

Remark: At this point, for the ease of presentation we focus on the case $\Psi(\boldsymbol{x}_i^1, \boldsymbol{x}_j^0) = \boldsymbol{x}_i^1 - \boldsymbol{x}_j^0$ (which is reflected in Algorithms 1 and 2). All the theory developed in this paper is valid for the most general case of $\Psi(\boldsymbol{x}_i^1, \boldsymbol{x}_j^0)$, however. In particular, imposing a norm bound on the individual features $\boldsymbol{x}_i^1, \boldsymbol{x}_j^0$ is equivalent to imposing a scaled norm bound on $\Psi(\boldsymbol{x}_i^1, \boldsymbol{x}_j^0)$.

Additional assumptions: We assume the input domain is compact, $\mathcal{X} \subseteq \mathcal{B}_2(X_*)$ which implies $\|\boldsymbol{x}\|_2 \leq X_*$, and that the domain of ranking functions is compact, taking $\mathcal{W} = \mathcal{B}_2(W_*)$ such that $\|\boldsymbol{w}\|_2 \leq W_*$. Here $\mathcal{B}_p(r) := \{\boldsymbol{x} \in \mathcal{X} : \|\boldsymbol{x}\|_p \leq r\}$ is the ℓ_p-ball of radius r. The first assumption is related to data preprocessing where the features are scaled appropriately; the second assumption corresponds to regularization. For both cases we chose the ℓ_2-norm as it provides a dimension-independent upper bound; however it is still possible to derive bounds using other norms. Let $\boldsymbol{w}_N$ and $\boldsymbol{w}_S$ be the minimizer of the empirical risk objectives (Algorithms 1 and 2, Line 10). Also for clarity we focus on the case where $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_S \succ \mathbf{0}$ in the sequel.[1] This means $\boldsymbol{w}_\star$, $\boldsymbol{w}_N$, and $\boldsymbol{w}_S$ have unique values. Note that, this last assumption is only for presentation purposes and the results derived in this paper apply to the most general case where $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_S \succeq \mathbf{0}$.

2 Batch Bipartite Ranking (BBR)

In order to analyze the more efficient LCBR, we first derive a risk bound for the corresponding BBR. The risk bounds derived in this paper are with respect to the best-in-class ranking function. For the sample set $\mathcal{N}$ we define $\rho := N_1/(N_0 + N_1)$ as the label skew. Also, for the unit sphere $\mathcal{S}^{D-1}$ let $C_{\mathcal{S}}(\epsilon)$ denote the covering number based on ℓ_2-balls of radius ϵ. The main result of this section is the following risk bound based on the metric entropy of $\mathcal{W}$.

[1] Where we defined $\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{x}_1 \sim P_1, \boldsymbol{x}_0 \sim P_0}[(\boldsymbol{x}_1 - \boldsymbol{x}_0)(\boldsymbol{x}_1 - \boldsymbol{x}_0)^\top]$.

Theorem 1. For a given sample set $\mathcal{N}$, define $\boldsymbol{w}_N := \arg\min_{\boldsymbol{w}\in\mathcal{W}} R_{\mathcal{N}}(\boldsymbol{w})$ and $\boldsymbol{w}_\star = \arg\min_{\boldsymbol{w}\in\mathcal{W}} R(\boldsymbol{w})$. Also define the constants $C_1 := 8X_*^2W_* + 4X_*$ and $C_2 := 3X_*^2W_*^2 + 2X_*W_*$. Then

$$\underset{\mathcal{N}\sim\mathcal{D}^N}{P}\left(R(\boldsymbol{w}_N) - R(\boldsymbol{w}_\star) \geq \epsilon\right) \leq 2\ C_{\mathcal{S}}\left(\frac{\epsilon}{4C_1}\right)\ \exp\left\{-\frac{\epsilon^2}{8C_2^2}\rho(1-\rho)N\right\}\ . \tag{4}$$

Here we used the shorthand $\mathcal{D}^N$ to denote the product measure over the samples, $[\mathcal{D}^1]^{N_1}\otimes[\mathcal{D}^0]^{N_0}$ where $[\mathcal{D}^i]^{N_i} = \otimes_{i=1}^{N_i}\mathcal{D}^i$. This result is comparable to the bound given for linear regression based on covering numbers [10]. One distinction, however, is the dependence on skew $\rho(1-\rho)$; when ρ is close to 0 or 1, the learner requires significantly more samples to achieve the same bound. The exponential term in Eq. (4) is comparable to the one obtained for AUC loss in Theorem 5 of [11]. On the other hand, when $\rho(1-\rho) = O(1)$ and N is large enough to bound the covering number by the exponential term in Eq. (4) we get the typical rate $O(\sqrt{\log(1/\delta)/N})$.

As the rate in Eq. (4) depends on $N_1 + N_0$ and not N_1N_0, it is natural to only consider pairs of independent samples instead of their Cartesian product, which reduces the analysis to that of linear regression. However, this is not done in practice as it would discard information [7]. Indeed, for this reason, a number of works consider the all-pair problem, e.g. [7, 4, 5]; we do so similarly. Below, we prove Theorem 1 using the following Lemma.

Lemma 1. For a given sample set $\mathcal{N}$ and constants $C_1 := 8X_*^2W_* + 4X_*$, $C_2 := 3X_*^2W_*^2 + 2X_*W_*$,

$$\underset{\mathcal{N}\sim\mathcal{D}^N}{P}\left(\sup_{\boldsymbol{w}\in\mathcal{W}}|R(\boldsymbol{w}) - R_{\mathcal{N}}(\boldsymbol{w})| \geq \epsilon\right) \leq 2\ C_{\mathcal{S}}\left(\frac{\epsilon}{2C_1}\right)\ \exp\left\{-\frac{\epsilon^2}{2C_2^2}\rho(1-\rho)N\right\}\ . \tag{5}$$

Proof. Recall that pairwise squared loss is defined as $\boldsymbol{\ell}(\boldsymbol{w}, \boldsymbol{x}^1, \boldsymbol{x}^0) = (1/2)(1 - h_{\boldsymbol{w}}(\boldsymbol{x}^1, \boldsymbol{x}^0))^2$ for $h_{\boldsymbol{w}}(\boldsymbol{x}^1, \boldsymbol{x}^0) = \boldsymbol{w}^\top(\boldsymbol{x}^1 - \boldsymbol{x}^0)$. Also define $\Phi_{\mathcal{N}}(\boldsymbol{w}) := R(\boldsymbol{w}) - R_{\mathcal{N}}(\boldsymbol{w})$. For $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}$,

$$\begin{aligned}
\left|\ell(\boldsymbol{w}_1, \boldsymbol{x}^1, \boldsymbol{x}^0) - \ell(\boldsymbol{w}_2, \boldsymbol{x}^1, \boldsymbol{x}^0)\right| &= \left|\frac{1}{2}(1 - h_{\boldsymbol{w}_1}(\boldsymbol{x}^1, \boldsymbol{x}^0))^2 - \frac{1}{2}(1 - h_{\boldsymbol{w}_2}(\boldsymbol{x}^1, \boldsymbol{x}^0))^2\right| \\
&\leq \frac{1}{2}\left|2 - h_{\boldsymbol{w}_1}(\boldsymbol{x}^1, \boldsymbol{x}^0) - h_{\boldsymbol{w}_2}(\boldsymbol{x}^1, \boldsymbol{x}^0)\right|\ \left|h_{\boldsymbol{w}_2}(\boldsymbol{x}^1, \boldsymbol{x}^0) - h_{\boldsymbol{w}_1}(\boldsymbol{x}^1, \boldsymbol{x}^0)\right| \\
&\leq (2X_*W_* + 1)\left|(\boldsymbol{w}_2 - \boldsymbol{w}_1)^\top(\boldsymbol{x}^1 - \boldsymbol{x}^0)\right| \\
&\leq (4X_*^2W_* + 2X_*)\left\|\boldsymbol{w}_1 - \boldsymbol{w}_2\right\|_2\ .
\end{aligned} \tag{6}$$

Using this bound we have

$$\begin{aligned}
&|\Phi_{\mathcal{N}}(\boldsymbol{w}_1) - \Phi_{\mathcal{N}}(\boldsymbol{w}_2)| = |R(\boldsymbol{w}_1) - R_{\mathcal{N}}(\boldsymbol{w}_1) - R(\boldsymbol{w}_2) + R_{\mathcal{N}}(\boldsymbol{w}_2)| \\
&\leq \left|\mathbb{E}_{\substack{x^1\sim D^1 \\ x^0\sim D^0}}\left[\ell(\boldsymbol{w}_1, \boldsymbol{x}^1, \boldsymbol{x}^0) - \ell(\boldsymbol{w}_2, \boldsymbol{x}^1, \boldsymbol{x}^0)\right]\right| + \frac{1}{N_1N_0}\sum_{i=1}^{N_1}\sum_{j=1}^{N_0}\left|\ell(\boldsymbol{w}_1, \boldsymbol{x}_i^1, \boldsymbol{x}_j^0) - \ell(\boldsymbol{w}_2, \boldsymbol{x}_i^1, \boldsymbol{x}_j^0)\right| \\
&\leq (8X_*^2W_* + 4X_*)\left\|\boldsymbol{w}_1 - \boldsymbol{w}_2\right\|_2.
\end{aligned} \tag{7}$$

We have shown that $|\Phi_{\mathcal{N}}(\boldsymbol{w}_1) - \Phi_{\mathcal{N}}(\boldsymbol{w}_2)| \leq C_1\left\|\boldsymbol{w}_1 - \boldsymbol{w}_2\right\|_2$. Let $\{\mathcal{B}_i\}_{i=1}^I$ be a set of ℓ_2-balls of

radius ϵ covering $\mathcal{W}$. Then

$$P_{\mathcal{N}}\left(\sup_{\boldsymbol{w}\in\mathcal{W}}|\Phi_{\mathcal{N}}(\boldsymbol{w})|\geq\epsilon\right)\leq\sum_{i=1}^{I}P_{\mathcal{N}}\left(\sup_{\boldsymbol{w}\in\mathcal{B}_i}|\Phi_{\mathcal{N}}(\boldsymbol{w})|\geq\epsilon\right)\leq\sum_{i=1}^{I}P_{\mathcal{N}}\left(|\Phi_{\mathcal{N}}(\boldsymbol{w}_i)|\geq\epsilon/2\right). \tag{8}$$

Now that the weight vector and samples are decoupled, we can bound the deviation of $\Phi_{\mathcal{N}}(\boldsymbol{w})$ for a fixed $\boldsymbol{w}$. First note that $\mathbb{E}[\Phi_{\mathcal{N}}(\boldsymbol{w})]=0$ by definition of $R(\boldsymbol{w})$ and $R_{\mathcal{N}}(\boldsymbol{w})$. Now consider perturbation of a single variable—define $\mathcal{N}'=(\boldsymbol{x}^1,\ldots,\boldsymbol{x}',\ldots,\boldsymbol{x}^N)$ which matches $\mathcal{N}$ everywhere except $\boldsymbol{x}'$. We have two cases:

(i) When $\boldsymbol{x}'$ has corresponding label 0, we have

$$\begin{aligned}|\Phi_{\mathcal{N}}(\boldsymbol{w})-\Phi_{\mathcal{N}'}(\boldsymbol{w})|&=\left|\frac{1}{2N_0N_1}\sum_{i=1}^{N_1}[1-\boldsymbol{w}^\top(\boldsymbol{x}_i^1-\boldsymbol{x}_j^0)]^2-\frac{1}{2N_0N_1}\sum_{i=1}^{N_1}[1-\boldsymbol{w}^\top(\boldsymbol{x}_i^1-\boldsymbol{x}')]^2\right|\\&\leq\frac{1}{2N_0N_1}N_1 6X_*^2W_*^2+\frac{1}{2N_0N_1}N_1 4X_*W_*\end{aligned} \tag{9}$$

From this last line it follows that $|\Phi_{\mathcal{N}}(\boldsymbol{w})-\Phi_{\mathcal{N}'}(\boldsymbol{w})|\leq C_2/N_0$.

(ii) Similarly when $\boldsymbol{x}'$ has corresponding label 1: $|\Phi_{\mathcal{N}}(\boldsymbol{w})-\Phi_{\mathcal{N}'}(\boldsymbol{w})|\leq C_2/N_1$.

As the differences are bounded and the inputs are independent, applying McDiarmid's inequality yields $P(|\Phi_{\mathcal{N}}(\boldsymbol{w})|\geq\epsilon/2)\leq 2\exp\{-(\epsilon^2/2C_2^2)\rho(1-\rho)N\}$. This, along with $I=C_{\mathcal{S}}(\epsilon/(2C_1))$, implies the bound in Lemma 1. □

Proof of Theorem 1. By definition of $\boldsymbol{w}_{\mathcal{N}}$ and $\boldsymbol{w}_\star$ we have $R(\boldsymbol{w}_N)-R(\boldsymbol{w}_\star)\geq 0$. Also,

$$\begin{aligned}R(\boldsymbol{w}_N)-R(\boldsymbol{w}_\star)&=[R(\boldsymbol{w}_N)-R_{\mathcal{N}}(\boldsymbol{w}_N)+R_{\mathcal{N}}(\boldsymbol{w}_\star)-R(\boldsymbol{w}_\star)]+[R_{\mathcal{N}}(\boldsymbol{w}_N)-R_{\mathcal{N}}(\boldsymbol{w}_\star)]\\&\leq|R(\boldsymbol{w}_N)-R_{\mathcal{N}}(\boldsymbol{w}_N)|+|R(\boldsymbol{w}_\star)-R_{\mathcal{N}}(\boldsymbol{w}_\star)|\ .\end{aligned} \tag{10}$$

where the second line follows from $R_{\mathcal{N}}(\boldsymbol{w}_N)-R_{\mathcal{N}}(\boldsymbol{w}_\star)\leq 0$ and the triangle inequality. We next bound both terms by $\epsilon/2$, and according to Lemma 1, the bounds hold simultaneously with probability at least $1-C_{\mathcal{S}}(\epsilon/4C_1)2\exp\{-(\epsilon^2/(8C_2^2))\rho(1-\rho)N\}$. □

3 Low Cost Bipartite Ranking (LCBR)

In this section we derive risk bounds for LCBR, a subsampling strategy for approximately optimizing BBR. Our main goal is to obtain a bound similar to that in Eq. (4). We start with a brief comparison of LBCR and BBR: In Section 1 we noted that the empirical risk objective of BBR can be written as $R_{\mathcal{N}}(\boldsymbol{w})=(1/2)\boldsymbol{w}^\top\boldsymbol{\Sigma}_N\boldsymbol{w}-\boldsymbol{\mu}_N^\top\boldsymbol{w}$ where $\boldsymbol{\Sigma}_N=1/(N_1N_0)\sum_{i=1}^{N_1}\sum_{j=1}^{N_0}(\boldsymbol{x}_i^1-\boldsymbol{x}_j^0)(\boldsymbol{x}_i^1-\boldsymbol{x}_j^0)^\top$ and $\boldsymbol{\mu}_N=1/(N_1N_0)\sum_{i=1}^{N_1}\sum_{j=1}^{N_0}(\boldsymbol{x}_i^1-\boldsymbol{x}_j^0)$. Here $\boldsymbol{\Sigma}_N$ and $\boldsymbol{\mu}_N$ are constructed using all N_1N_0 pairs available. On the other hand, LBCR subsamples S *pairs* from the fixed

$N_1 N_0$ *total pairs* uniformly at random with replacement. The pairs obtained this way are denoted by the set $\mathcal{S} = \{(x_1^1, x_1^0), \ldots, (x_S^1, x_S^0)\}$. It can be seen that the elements of $\mathcal{S}$ are random variables sampled from a uniform distribution conditional on $\mathcal{N}$. Thus, while the elements of $\mathcal{N}$ are sampled from the class-conditionals, i.e. $\mathcal{D}^i$, the elements of $\mathcal{S}$ are sampled from the uniform distribution $\mathcal{D}(\mathcal{N})$. With a slight abuse of notation, we will denote this by $(\boldsymbol{x}_i^1, \boldsymbol{x}_j^0) \sim \mathcal{N}$. The corresponding objective of LCBR is $R_{\mathcal{S}}(\boldsymbol{w}) = (1/2)\boldsymbol{w}^\top \boldsymbol{\Sigma}_S \boldsymbol{w} - \boldsymbol{\mu}_S^\top \boldsymbol{w}$ with $\boldsymbol{\mu}_S$ and $\boldsymbol{\Sigma}_S$ the first and second moments computed on the subsample.

In order to derive a risk bound for $|R(\boldsymbol{w}_S) - R(\boldsymbol{w}_\star)|$ we first need to bound $|R_{\mathcal{N}}(\boldsymbol{w}_N) - R_{\mathcal{N}}(\boldsymbol{w}_S)|$. However, in the latter expression, the weight vectors and the samples are not independent. For this reason, we again use a uniform convergence argument. Here we use matrix and vector concentration to obtain the bounds necessary. In particular, the following lemma provides concentration inequalities for two key variables.

Lemma 2. For $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}$, define $\Delta_{\boldsymbol{\Sigma}} := \boldsymbol{w}_1^\top (\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N)\boldsymbol{w}_1$ and $\Delta_\sigma := (\boldsymbol{w}_1 - \boldsymbol{w}_2)^\top (\boldsymbol{\mu}_N - \boldsymbol{\mu}_S)$. The following hold:

(i) $P(\sup_{\boldsymbol{w}_1 \in \mathcal{W}} |\Delta_{\boldsymbol{\Sigma}}| \geq \epsilon) \leq 2D \exp\{-S\epsilon^2 / (8 \left\|\boldsymbol{\Sigma}_N\right\|_2 X_*^2 W_*^4 + (16/3)\epsilon X_*^2 W_*^2)\}$

(ii) $P(\sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}} |\Delta_\sigma| \geq \epsilon) \leq 2 \exp\{-1/2(\epsilon\sqrt{S}/(2X_* W_*) - 1)^2\}$

The proof is given in the appendix. We can now bound the difference of the empirical risks with high probability.

Theorem 2. For a given sample set $\mathcal{N}$ and its subsample set $\mathcal{S}$ let $\boldsymbol{w}_N = \arg\min_{\boldsymbol{w} \in \mathcal{W}} R_{\mathcal{N}}(\boldsymbol{w})$ and $\boldsymbol{w}_S = \arg\min_{\boldsymbol{w} \in \mathcal{W}} R_{\mathcal{S}}(\boldsymbol{w})$. If the subsample size satisfies

$$S \geq \max\left\{ \log(4D/\delta) \frac{\left\|\boldsymbol{\Sigma}_N\right\|_2 X_*^2 W_*^4 + (1/3)\epsilon X_*^2 W_*^2}{\epsilon^2/32} \;,\; \frac{X_*^2 W_*^2}{\epsilon^2/4} \left[\sqrt{2\log(4/\delta)} + 1\right]^2 \right\} \tag{11}$$

then

$$\underset{\mathcal{S} \sim \mathcal{N}}{P}\left(|R_{\mathcal{N}}(\boldsymbol{w}_S) - R_{\mathcal{N}}(\boldsymbol{w}_N)| \geq \epsilon \right) \leq 1 - \delta \;. \tag{12}$$

Proof. We define the following pointwise difference

$$\Delta(\boldsymbol{w}) := R_{\mathcal{S}}(\boldsymbol{w}) - R_{\mathcal{N}}(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top (\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N)\boldsymbol{w} + \boldsymbol{w}^\top (\boldsymbol{\mu}_N - \boldsymbol{\mu}_S) \;. \tag{13}$$

Our first task is to show that the following inequality holds,

$$0 \leq R_{\mathcal{N}}(\boldsymbol{w}_S) - R_{\mathcal{N}}(\boldsymbol{w}_N) \leq \Delta(\boldsymbol{w}_N) - \Delta(\boldsymbol{w}_S) \;. \tag{14}$$

The LHS of this inequality holds by the definition of $\boldsymbol{w}_N$ and $\boldsymbol{w}_S$. For the RHS we have

$$R_{\mathcal{N}}(\boldsymbol{w}_S) - R_{\mathcal{N}}(\boldsymbol{w}_N) = \Delta(\boldsymbol{w}_N) - \Delta(\boldsymbol{w}_S) - [R_{\mathcal{S}}(\boldsymbol{w}_N) - R_{\mathcal{S}}(\boldsymbol{w}_S)] \leq \Delta(\boldsymbol{w}_N) - \Delta(\boldsymbol{w}_S) \tag{15}$$

since $R_{\mathcal{S}}(\boldsymbol{w}_N) - R_{\mathcal{S}}(\boldsymbol{w}_S) \geq 0$. It is therefore sufficient prove a high probability bound for $|\Delta(\boldsymbol{w}_N) - \Delta(\boldsymbol{w}_S)|$. Next,

$$\begin{aligned}|\Delta(\boldsymbol{w}_N) - \Delta(\boldsymbol{w}_S)| &= \left|\frac{1}{2}\boldsymbol{w}_N^\top(\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N)\boldsymbol{w}_N - \frac{1}{2}\boldsymbol{w}_S^\top(\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N)\boldsymbol{w}_S + (\boldsymbol{w}_S - \boldsymbol{w}_N)^\top(\boldsymbol{\mu}_N - \boldsymbol{\mu}_S)\right| \\ &\leq \left|\frac{1}{2}\boldsymbol{w}_N^\top(\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N)\boldsymbol{w}_N - \frac{1}{2}\boldsymbol{w}_S^\top(\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N)\boldsymbol{w}_S\right| + \left|(\boldsymbol{w}_S - \boldsymbol{w}_N)^\top(\boldsymbol{\mu}_N - \boldsymbol{\mu}_S)\right|. \quad (16)\end{aligned}$$

We bound each of these terms by $\epsilon/2$ with probability at least $1 - \delta/2$. Note that both terms have weight vectors coupled with samples so we cannot apply concentration inequalities directly. However, they can be upper bounded by the expressions in Lemma 2. In particular, for the quadratic term $\left|\frac{1}{2}\boldsymbol{w}_N^\top(\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N)\boldsymbol{w}_N - \frac{1}{2}\boldsymbol{w}_S^\top(\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N)\boldsymbol{w}_S\right| \leq \sup_{\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathcal{W}}|\Delta_{\boldsymbol{\Sigma}}|$. We then apply Lemma 5(i) with threshold $\epsilon/2$ and probability $\delta/2$ which yields the first term in Eq. (11). For the linear term, $\left|(\boldsymbol{w}_S - \boldsymbol{w}_N)^\top(\boldsymbol{\mu}_N - \boldsymbol{\mu}_S)\right| \leq \sup_{\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathcal{W}}|\Delta_\sigma|$. Applying Lemma 5(ii) with threshold $\epsilon/2$ and probability $\delta/2$ yields the second term. □

Theorem 2 shows that the subsample size S required to decrease the difference $|R_{\mathcal{N}}(\boldsymbol{w}_S) - R_{\mathcal{N}}(\boldsymbol{w}_N)|$ does not depend on the number of total pairs N_1N_0. Instead, the dependence is on the operator norm of the empirical second moment matrix $\|\boldsymbol{\Sigma}_N\|_2$, and polynomial in X_*, W_*, $\log(1/\delta)$, and $1/\epsilon$. This is favorable, as in many settings the total number of pairs can be prohibitively large. The following is the main result of this section, regarding the actual risk of LCBR solution.

Theorem 3. For a probability target $p^\star$ let the sample size be chosen such that

$$S \geq \max\left\{\log(4D/p^\star)\frac{\|\boldsymbol{\Sigma}_N\|_2 X_*^2W_*^4 + (1/15)\epsilon X_*^2W_*^2}{\epsilon^2/800}\ ,\ \frac{X_*^2W_*^2}{\epsilon^2/100}\left[\sqrt{2\log(4/p^\star)}+1\right]^2\right\}. \quad (17)$$

Then the solution $\boldsymbol{w}_S$ returned by LCBR satisfies

$$\underset{\mathcal{N}\sim\mathcal{D}^N}{P}\left(R(\boldsymbol{w}_S) - R(\boldsymbol{w}_\star) \geq \epsilon\right) \leq 2\ C_{\mathcal{S}}\left(\frac{\epsilon}{10C_1}\right)\ \exp\left\{-\frac{\epsilon^2}{50C_2^2}\rho(1-\rho)N\right\} + p^\star\ . \quad (18)$$

Proof. We start with the inequality

$$R(\boldsymbol{w}_S) - R(\boldsymbol{w}_\star) = |R(\boldsymbol{w}_S) - R(\boldsymbol{w}_\star)| \leq |R(\boldsymbol{w}_S) - R(\boldsymbol{w}_N)| + |R(\boldsymbol{w}_N) - R(\boldsymbol{w}_\star)|\,. \quad (19)$$

The first term can be bounded as

$$\begin{aligned}|R(\boldsymbol{w}_S) - R(\boldsymbol{w}_N)| &= |R(\boldsymbol{w}_S) - R_{\mathcal{N}}(\boldsymbol{w}_S) + R_{\mathcal{N}}(\boldsymbol{w}_S) - R_{\mathcal{N}}(\boldsymbol{w}_N) + R_{\mathcal{N}}(\boldsymbol{w}_N) - R(\boldsymbol{w}_N)| \\ &\leq |R(\boldsymbol{w}_S) - R_{\mathcal{N}}(\boldsymbol{w}_S)| + |R(\boldsymbol{w}_N) - R_{\mathcal{N}}(\boldsymbol{w}_N)| + |R_{\mathcal{N}}(\boldsymbol{w}_S) - R_{\mathcal{N}}(\boldsymbol{w}_N)|\ .\end{aligned} \quad (20)$$

Note that the empirical risk trick in the proof Theorem 1 no longer applies as it not known which one of $R(\boldsymbol{w}_N)$ or $R(\boldsymbol{w}_S)$ is smaller, and all three terms have to be retained. On the other hand, for the second term, from the proof of Theorem 1 we know that

$$|R(\boldsymbol{w}_N) - R(\boldsymbol{w}_\star)| \leq |R(\boldsymbol{w}_N) - R_{\mathcal{N}}(\boldsymbol{w}_N)| + |R(\boldsymbol{w}_\star) - R_{\mathcal{N}}(\boldsymbol{w}_\star)|\ . \quad (21)$$

Combining Eqs. (20) and (21) we get

$$R(\boldsymbol{w}_S) - R(\boldsymbol{w}_\star) \leq \left[\sum_{\boldsymbol{w} \in \mathcal{W}_4} |R(\boldsymbol{w}) - R_{\mathcal{N}}(\boldsymbol{w})| \right] + |R_{\mathcal{N}}(\boldsymbol{w}_S) - R_{\mathcal{N}}(\boldsymbol{w}_N)| \ . \tag{22}$$

where we defined the sequence $\mathcal{W}_4 = [\boldsymbol{w}_\star, \boldsymbol{w}_N, \boldsymbol{w}_N, \boldsymbol{w}_S]$. We now consider bounding each term by $\epsilon/5$. For the summation on the right hand side this yields

$$\begin{aligned} P\left(\sum_{\boldsymbol{w} \in \mathcal{W}_4} |R(\boldsymbol{w}) - R_{\mathcal{N}}(\boldsymbol{w})| \leq 4\epsilon/5 \right) &\leq P\left(\sup_{\boldsymbol{w} \in \mathcal{W}} |R(\boldsymbol{w}) - R_{\mathcal{N}}(\boldsymbol{w})| \leq \epsilon/5 \right) \\ &\leq 2\ C_{\mathcal{S}}\left(\frac{\epsilon}{10 C_1} \right)\ \exp\left\{ -\frac{\epsilon^2}{50 C_2^2} \rho(1-\rho) N \right\} \end{aligned} \tag{23}$$

where the last inequality follows from Lemma 1. For the last term we would like to bound the deviation by $\epsilon/5$ with probability $p^\star$. Plugging these terms into the sample complexity bound of Theorem 2 yields the bound of Eq. (17). □

Theorem 3 shows that the number of samples S required to make LCBR competitive with BBR does not depend on $N_1 N_0$. Firstly, for a fixed probability target $p^\star$, S depends on $\mathcal{N}$ only through $\boldsymbol{\Sigma}_N$. In practice we set $p^\star$ such that it is relatively smaller than the exponential term in Eq. (18). In this case S will have an implicit dependence on $N_1 + N_0$ and ρ. The important difference is, while BBR requires $N_1 N_0$ samples to achieve the bound in Eq. (4), LCBR can achieve the comparable bound in Eq. (18) with $S \ll N_1 N_0$. We demonstrate this with experiments in Section 5. Finally, note that we focused on the linear case due to space limitations. Clearly, for any *fixed* nonlinear feature map, our results hold where D is replaced with the dimensionality of the new feature space. Another direction is to consider random feature transforms and provide bounds based to the best ranker in the corresponding function space [12]; we leave this for a longer version of the paper.

4 Related Work

Our proof of the uniform risk bound in Lemma 1 is based on covering numbers, which is also used for analyzing linear regression problems [10]. The covering number-based argument is later extended to online learning with pairwise loss functions [13]; however, their analysis is for sequential updates. Bounds based on Rademacher complexity and U-processes are considered in [9, 14, 15, 16]. These bounds can be tighter, however they are based on the assumption that all samples are drawn i.i.d. from $\mathcal{D}$, which is different from our setup. Bounds based on algorithmic stability [17, 18] and VC dimenstion [11] have also been considered. Replacing the discrete AUC loss with a convex one has been investigated in a number of work. In particular [4, 5, 9, 8, 7] consider online algorithms. Online learning based on stochastic saddle point problems is considered in [19, 20]. It is also worthwhile to note that the majority of the aforementioned papers are based on the pairwise squared loss. In [6], the consistency of surrogate loss functions with respect to AUC loss has been considered, extending the results of [21]. On the other hand

[22, 23] provide bounds for the AUC loss in terms of the surrogate loss of the learner. Kernel based methods for bipartite ranking was considered in [3, 24, 25]. (For more related work see also [26, 27, 28].) In terms of low sample complexity, the stochastic gradient descent (SGD) based online learning algorithms also provide bounds in terms of the samples used; however these algorithms require a step size to tune, whereas LCBR in Algorithm 2 does not need this parameter, which can be an important practical advantage. As we will show in experiments, LCBR can achieve better performance with fewer samples, compared to the SGD-based methods.

5 Experiments

5.1 Gaussian Mixture Distribution

We first consider experiments where the generating distribution is a K-component Gaussian Mixture. So for $i \in \{0, 1\}$ we can write

$$\mathcal{D}^i(\boldsymbol{x}) = \sum_{k=1}^{K} \frac{c_k}{\sqrt{2\pi\sigma^2}} \exp\{-\left\|\boldsymbol{x} - \boldsymbol{\mu}_k^i\right\|_2^2 / (2\sigma^2)\} \tag{24}$$

where c_k's are mixture weights. We consider the case where the covariance matrix is isotropic and controlled with a single scale parameter σ. We consider three cases where $K = 1, 2, 3$ and $\sigma = 2, 3, 4$ where the increasing value of K and σ makes the problem gradually more difficult. The mean values for class one are sampled from the unit cube in positive orthant and class zero from unit cube in negative orthant. The curves are obtained by averaging 50 experiments. For any given pair of conditional distributions the optimal ranking function minimizing the ϕ-risk is found by calculating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and solving the ϕ-risk minimization problem. On the other hand, the optimal ranking function maximizing the AUC is the likelihood ratio of class-conditionals, which is a consequence of the Neyman-Pearson lemma [29]. Figure 1 shows the results of our experiments. For panels (a)-(f) we have $\rho = 0.5$ with 10^3 positive and negative samples. So BBR uses 10^6 pairs whereas LCBR uses the number shown on the x-axis. In panels (a)-(c) we show the ϕ-risk and in (d)-(f) we show the AUC. As the number of components increase the problem becomes more difficult, resulting in higher ϕ-risk and lower AUC for optimal ranking functions. In all cases BBR is very close to the optimal ranker; but it has high sample cost. LCBR, on the other hand, catches on with $S = 3000$ at most. This is also reflected in wall clock times; as shown in panels (g)-(i) LCBR is roughly 100 times faster than BBR when $S = 5000$ subsamples are used. Finally we illustrate the performance of BBR and LCBR as a function of label skew ρ. In the proof of Lemma 1, we saw that applying McDiarmid's inequality gives the term $\rho(1-\rho)$ in the exponent. This suggests, as ρ goes to 0 or 1 the generalization performance should decrease. Panel (j) shows that this is indeed the case: here BBR uses 2000ρ positive and $2000(1-\rho)$ negative samples, whereas LCBR uses $5,000$ subsamples. The performance of BBR and LCBR are once again close; but as ρ deviates from 0.5 the generalization of both algorithms get worse.

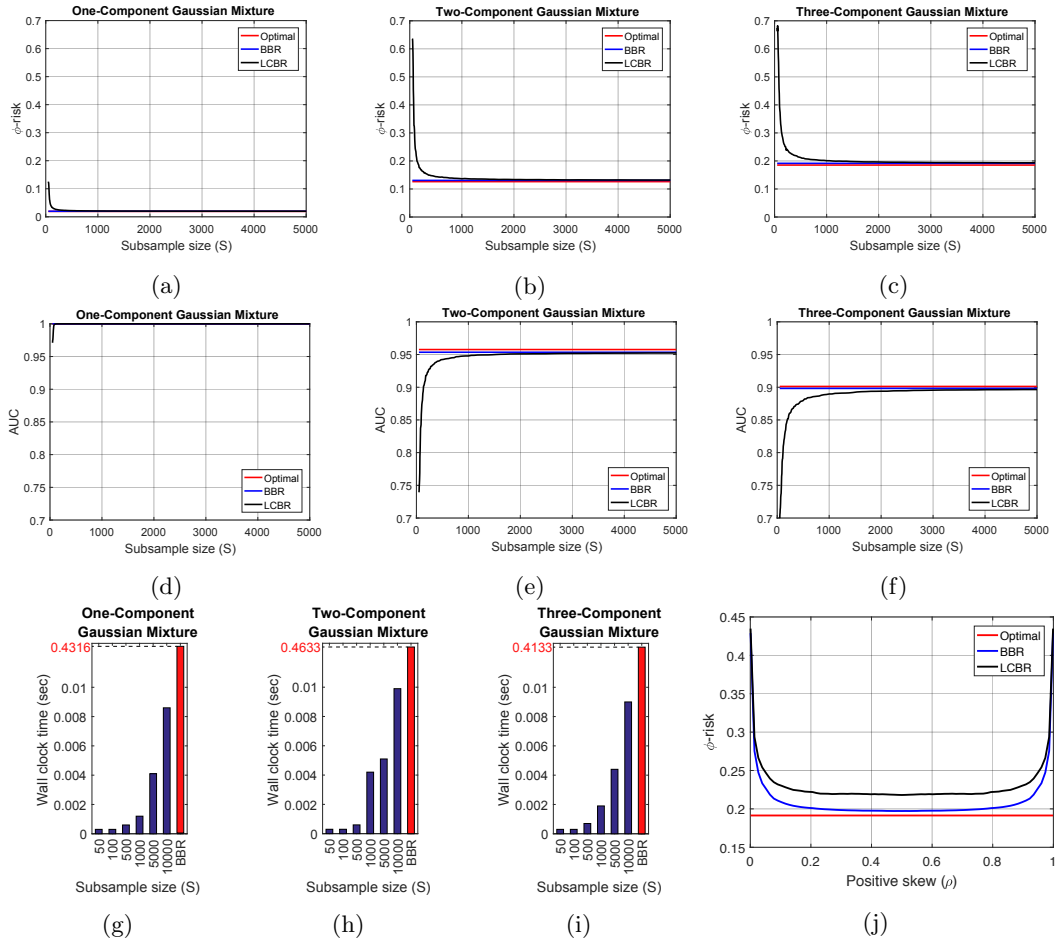
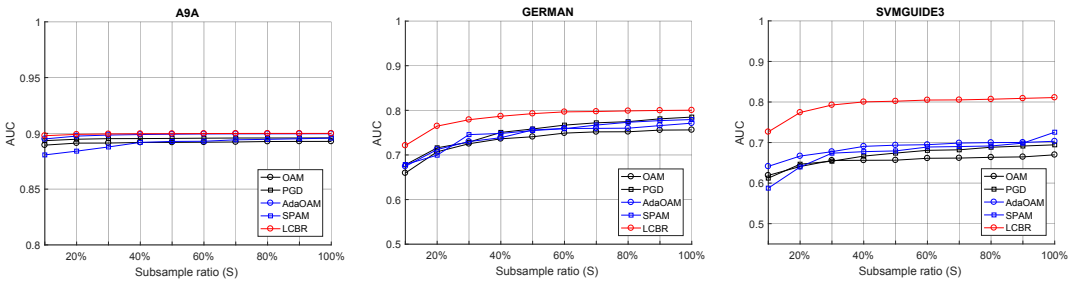


Figure 1: Results of experiments with Gaussian Mixture distributions.

5.2 LIBSVM Datasets

We now compare LCBR against state-of-the art algorithms on three datasets from LIBSVM. The algorithms we implement, in addition to LCBR, are the following: AdaOAM [8] uses adaptive stochastic gradient descent with pairwise squared loss, whereas PGD is SGD based approach [7]. SPAM is one of the most recent works where AUC optimization is formulated as a stochastic saddle point problem [19]. OAM is a relatively older algorithm, but we include it as it is based in a different objective, the pairwise hinge loss [4]. As widely done in the literature, we use AUC on test set as the performance measure. We use the train and test splits provided by LIBSVM. Regularization parameters are determined by cross-validation and step sizes are chosen based on the references. The experiments are averaged over 50 runs.



(a) (b) (c)

Figure 2: Results of experiments with LIBSVM datasets.

We show the results in Figure 2. The x-axis correspond to the number of subsamples used as a percentage of total number of samples available. For example, `A9A` dataset contains approximately $33K$ data points. A subsample ratio of 50% implies we use approximately $16.5K$ random samples from this dataset. For the `A9A` dataset all algorithms have good generalization, with an AUC of approximately 90%. For the `GERMAN` and `SVMGUIDE3` datasets, on the other hand, there is a significant gap between LCBR and the SGD-based competitiors. Here the SGD and learning rate free approach of LCBR proves useful and it achieves better generalization for the given number of samples. Therefore the sample complexity of LCBR is favorable.

5.3 Algorithmic Complexity

For the algorithms presented in this paper, it can be seen that the computational bottleneck is at the accumulation phase (cf. Algorithms 1 and 2). For BBR the cost of this step is $O(N_1 N_0 D^2)$ and for LCBR this is $O(SD^2)$. Typically $S \ll N_1 N_0$ which can save significant computation. On the other hand, the storage for both BBR and LCBR is $O(D^2)$ as it requires storing the covariance matrix. This quadratic dependence on dimension can be an issue when D is too large. In this case, a sparse approximation of $\mathbf{\Sigma}_S$ can be necessary. However, note that this storage bottleneck is also present in the algorithms we compared [4, 7, 8]. In contrast, SPAM [19] can be implemented in $O(D)$ space; however that algorithm requires knowledge of expectations with respect to the true conditionals, which is unknown in practice. Our experiments show that LCBR can still achieve better performance.

6 Conclusion

We have considered the problem of bipartite ranking, where the empirical AUC loss is replaced with the pairwise squared loss. Different from the previous work—which was based on SGD—we proposed a low sample cost bipartite ranking algorithm (LCBR), and showed that the number

of samples required for good performance obeys $S \ll N_1 N_0$. Experiments show that LCBR quickly achieves similar performance with BBR where the number of samples are several order of magnitudes lower. Experiments against state-of-the-art bipartite ranking algorithms also show that LCBR can achieve better generalization with a smaller subsample set. In a longer version of the paper we will also consider extending these results to random feature spaces, which include random kernel features and random neural networks.

7 Appendix: Proof of Lemma 2

For the proofs we will use the shorthand $\boldsymbol{x}_s = \boldsymbol{x}^1_{i_s} - \boldsymbol{x}^0_{j_s}$.

(i) We recall the Matrix Bernsten Inequality for a $D \times D$ symmetric, random matrix $\boldsymbol{Z}$ and threshold γ:

$$P(\|\boldsymbol{Z}\|_2 > \gamma) \leq 2D \exp\left(-\frac{\gamma^2/2}{\mathbb{V}(\boldsymbol{Z}) + L\gamma/3}\right) \tag{25}$$

where L is a norm bound on summands.

First apply a spectral norm bound to the quadratic expression: $\sup_{\boldsymbol{w}_1 \in \mathcal{W}} |\Delta_{\boldsymbol{\Sigma}}| \leq W_*^2 \|\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N\|_2$. We now take $\boldsymbol{Z} = \boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_N$. The spectral norm can be bounded based on the argument in [30]. We can decompose $\boldsymbol{Z}$ into a sum: $\boldsymbol{Z} = \sum_{s=1}^{S} (1/S)[\boldsymbol{x}_s \boldsymbol{x}_s^\top - \boldsymbol{\Sigma}_N]$. We denote each summand by $\boldsymbol{E}_s = (1/S)[\boldsymbol{x}_s \boldsymbol{x}_s^\top - \boldsymbol{\Sigma}_N]$. It then follows from triangle inequality that

$$\|\boldsymbol{E}_s\|_2 \leq \frac{1}{S}\left[\left\|\boldsymbol{x}_s \boldsymbol{x}_s^\top\right\|_2 + \|\boldsymbol{\Sigma}_N\|_2\right] \leq \frac{2}{S}\|\boldsymbol{x}_s\|_2^2 \leq \frac{8X_*^2}{S} \tag{26}$$

where the second inequality follows from Jensen's inequality. Since the subsampling is conditional on $\mathcal{N}$, $\boldsymbol{\Sigma}_N$ is constant with respect to the subsampled pairs, and each summand is centered and i.i.d. The variance of the sum decomposes as $\mathbb{V}(\boldsymbol{Z}) = \left\|\sum_{s \in S} \mathbb{E}[\boldsymbol{E}_s^2]\right\|_2$. For a single summand the second moment can be bounded as

$$\mathbb{E}[\boldsymbol{E}_s^2] = \frac{1}{S^2}\mathbb{E}\left[\left[\boldsymbol{x}_s \boldsymbol{x}_s^\top - \boldsymbol{\Sigma}_N\right]^2\right] = \frac{1}{S^2}\left[\mathbb{E}\left[\|\boldsymbol{x}_s\|_2^2 \boldsymbol{x}_s \boldsymbol{x}_s^\top\right] - \boldsymbol{\Sigma}_N^2\right] \preceq \frac{4X_*^2}{S^2}\boldsymbol{\Sigma}_N \tag{27}$$

from which the variance inequality $\mathbb{V}(\boldsymbol{Z}) \leq (4X_*^2/S)\|\boldsymbol{\Sigma}_N\|_2$ follows. Substituting these to Eq. (25) with $\gamma = \epsilon/W_*^2$ yields the result.

(ii) We recall the following concentration inequality for i.i.d. and bounded random vectors with mean $\bar{\boldsymbol{x}}$ [12]:

$$P\left(\left\|\frac{1}{S}\sum_{s=1}^{S}\boldsymbol{x}_s - \bar{\boldsymbol{x}}\right\|_2 \geq \gamma\right) \leq \exp\left\{-\frac{1}{2}\left(\frac{\gamma\sqrt{S}}{L} - 1\right)^2\right\} . \tag{28}$$

From the following inequalities

$$\sup_{\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathcal{W}} |\Delta_\sigma(\boldsymbol{w}_1,\boldsymbol{w}_2)| \leq \sup_{\boldsymbol{w}_1,\boldsymbol{w}_2} \|\boldsymbol{w}_1-\boldsymbol{w}_2\|_S \|\boldsymbol{\mu}_N-\boldsymbol{\mu}_S\|_2 \leq 2W_* \left\| \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{x}_s - \boldsymbol{\mu}_N \right\|_2 \tag{29}$$

the desired result is obtained by setting $\gamma = \epsilon/2W_*$ and $L = X_*$ in Eq. (28). □

References

[1] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica, "Ad click prediction: a view from the trenches," in *Conference on Knowledge Discovery and Data Mining*, 2013.

[2] J. Hu, H. Yang, M. Lyu, I. King, and A. So, "Online Nonlinear AUC Maximization for Imbalanced Data Sets," *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[3] X. Hong, S. Chen, and C. J. Harris, "A Kernel-Based Two-Class Classifier for Imbalanced Data Sets," *IEEE Transactions on Neural Networks*, 2007.

[4] P. Zhao, S. C. H. Hoi, R. Jin, and T. Yang, "Online AUC Maximization," in *International Conference on Machine Learning*, 2011.

[5] W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou, "One-Pass AUC Optimization," in *International Conference on Machine Learning*, 2013.

[6] W. Gao and Z.-H. Zhou, "On the Consistency of AUC Pairwise Optimization," in *International Joint Conference on Artificial Intelligence*, 2015.

[7] M. Boissier, S. Lyu, Y. Ying, and D.-X. Zhou, "Fast Convergence of Online Pairwise Learning Algorithms," in *International Conference on Artificial Intelligence and Statistics*, 2016.

[8] Y. Ding, P. Zhao, S. C. H. Hoi, and Y. S. Ong, "An Adaptive Gradient Method for Online AUC Maximization," in *Association for Advancement of Artificial Intelligence*, 2015.

[9] P. Kar, B. Sriperumbudur, P. Jain, and H. Karnick, "On the Generalization Ability of Online Learning Algorithms for Pairwise Loss Functions," in *International Conference on Machine Learning*, 2013.

[10] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT Press, 2012.

[11] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth, "Generalization Bounds for the Area Under the ROC Curve," *Journal of Machine Learning Research*, 2005.

[12] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Advances in Neural Information Processing Systems*, 2008.

[13] Y. Wang, R. Khardon, D. Pechyony, and R. Jones, "Generalization Bounds for Online Learning Algorithms with Pairwise Loss Functions," in *Conference on Learning Theory*, 2012.

[14] S. Clemencon, G. Lugosi, and N. Vayatis, "Ranking and empirical minimization of U-Statistics," *The Annals of Statistics*, 2008.

[15] T. Peel, A. Sandrine, and L. Ralaivola, "Empirical Bernstein Inequalities for U-Statistics," in *Advances in Neural Information Processing Systems*, 2010.

[16] Y. Lei, S.-B. Lin, and K. Tang, "Generalization Bounds for Regularized Pairwise Learning," in *International Joint Conference on Artificial Intelligence*, 2018.

[17] S. Agarwal and P. Niyogi, "Generalization Bounds for Ranking Algorithms via Algorithmic Stability," *Journal of Machine Learning Research*, 2009.

[18] C. Cortes, M. Mohri, and A. Rastogi, "Magnitude-Preserving Ranking Algorithms," in *International Conference on Machine Learning*, 2007.

[19] M. Natole Jr., Y. Ying, and S. Lyu, "Stochastic Proximal Algorithms for AUC Maximization," in *International Conference on Machine Learning*, 2018.

[20] Y. Ying, L. Wen, and S. Lyu, "Stochastic Online AUC Maximization," in *Advances in Neural Information Processing Systems*, 2016.

[21] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, 2006.

[22] S. Agarwal, "Surrogate Regret Bounds for Bipartite Ranking via Strongly Proper Losses," *Journal of Machine Learning Research*, 2014.

[23] W. Kotlowski, K. J. Dembczynski, and E. Huellermeier, "Bipartite ranking through minimization of univariate loss," in *International Conference on Machine Learning*, 2011.

[24] Y. Ding, C. Liu, P. Zhao, and S. C. Hoi, "Large Scale Kernel Methods for Online AUC Maximization," in *International Conference on Data Mining (ICDM)*, 2017.

[25] J. Hu, H. Yang, M. Lyu, I. King, and A. So, "Kernelized Online Imbalanced Learning with Fixed Budgets," in *Association for Advancement of Artificial Intelligence (AAAI)*, 2015.

[26] H. Narasimhan and S. Agarwal, "On the Relationship Between Binary Classification, Bipartite Ranking, and Binary Class Probability Estimation," in *Advances in Neural Information Processing Systems*, 2013.

[27] S. Agarwal, "Generalization bounds for some ordinal regression algorithms," in *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, 2008.

[28] A. Rajkumar and S. Agarwal, "When Can We Rank Well from Comparisons of O(n log n) Non-Actively Chosen Pairs?" in *Conference on Learning Theory*, 2016.

[29] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer, 1998.

[30] J. Tropp, "An introduction to matrix concentration inequalities," *Foundations and Trends in Machine Learning*, 2015.