Generalization Bounds for Online Learning Algorithms with Pairwise Loss Functions

Yuyang Wang YWANG02@CS.TUFTS.EDU
Roni Khardon RONI@CS.TUFTS.EDU
Department of Computer Science, Tufts University, Medford, MA 02155, USA

Dmitry Pechyony DPECHYON@AKAMAI.COM
Rosie Jones REJONES@AKAMAI.COM
Akamai Technologies, 8 Cambridge Center, Cambridge, MA 02142, USA

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

Efficient online learning with pairwise loss functions is a crucial component in building large-scale learning system that maximizes the area under the Receiver Operator Characteristic (ROC) curve. In this paper we investigate the generalization performance of online learning algorithms with pairwise loss functions. We show that the existing proof techniques for generalization bounds of online algorithms with a pointwise loss can not be directly applied to pairwise losses. Using the Hoeffding-Azuma inequality and various proof techniques for the risk bounds in the batch learning, we derive data-dependent bounds for the average risk of the sequence of hypotheses generated by an *arbitrary* online learner in terms of an easily computable statistic, and show how to extract a low risk hypothesis from the sequence. In addition, we analyze a natural extension of the perceptron algorithm for the bipartite ranking problem providing a bound on the empirical pairwise loss. Combining these results we get a complete risk analysis of the proposed algorithm.

Keywords: Generalization bounds, Pairwise loss functions, Online learning, Loss bounds.

1. Introduction

The standard framework in learning theory considers learning from examples $Z^n = \{(\boldsymbol{x}_t, y_t) \in \mathcal{X} \times \mathcal{Y}\},\ t = 1, 2, \cdots, n$, drawn at random from an unknown probability distribution $\mathcal{D}$ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ (e.g. $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$). Typically the loss associated with prediction is based on a single example and expressed as $\ell(h, (\boldsymbol{x}, y))$. In this paper we study learning in the context of pairwise loss functions, that depend on pairs of examples and can be expressed as $\ell\left(h, (\boldsymbol{x}, y), (\boldsymbol{u}, v)\right)$. Pairwise loss functions capture ranking problems that are important for a wide range of applications. For example, in the *supervised ranking* problem one wishes to learn a ranking function that predicts the correct ordering of objects. The *misranking loss* (Clemençon et al., 2008; Peel et al., 2010) is a pairwise loss such that $\ell_{\text{rank}}\left(h, (\boldsymbol{x}, y), (\boldsymbol{u}, v)\right) = \mathbb{I}_{[(y-v)(h(\boldsymbol{x})-h(\boldsymbol{u}))<0]}$, where $\mathbb{I}$ is the indicator function and the loss is 1 when the examples are ranked in the wrong order. The goal of learning is to find a hypothesis h that minimizes the expected misranking risk $\mathcal{R}(h)$,

$$\mathcal{R}(h) := \mathbb{E}_{(\boldsymbol{x},y)}\mathbb{E}_{(\boldsymbol{u},v)}\left[\ell\left(h, (\boldsymbol{x}, y), (\boldsymbol{u}, v)\right)\right]. \tag{1}$$

This problem, especially the bipartite ranking problem where $\mathcal{Y} = \{+1, -1\}$, has been extensively studied over the past decade in the *batch setting*, i.e., where the entire sequence Z^n is presented to the learner in advance of learning. On the empirical end, many algorithms have been proposed and successfully applied, for example, AUC Support Vector Machine (SVM) (Brefeld and Scheffer, 2005), Ranking SVM (Joachims, 2002), and RankBoost (Freund et al., 2003). Several theoretical studies also investigated the batch setting, deriving risk bounds for specific algorithms (Freund et al., 2003; Rudin et al., 2005), and uniform convergence bounds for empirical estimates of the risk (Agarwal et al., 2005; Agarwal and Niyogi, 2005, 2009) and related U-statistics (Clemençon et al., 2008; Peel et al., 2010).

In this paper we investigate the generalization performance of *online learning* algorithms, where examples are presented in sequence, in the context of pairwise loss functions. Specifically, on each round t, an online learner receives an instance $\boldsymbol{x}_t$ and predicts a label $\hat{y}_t$ according to the current hypothesis h_{t-1}. The true label y_t is revealed and h_{t-1} is updated. The goal of the online learner is to minimize the expected risk w.r.t. a pairwise loss function ℓ. Online learning algorithms have been studied extensively, and theoretical results provide relative loss bounds, where the online learner competes against the best hypothesis (with hindsight) on the same sequence. Conversions of online learning algorithms and their performance guarantees to provide generalization performance in the batch setting have also been investigated (e.g., (Kearns et al., 1987; Littlestone, 1990; Freund and Schapire, 1999; Zhang, 2005)). Cesa-Bianchi et al. (2004) provided a general online-to-batch conversion result that holds under some mild assumptions on the loss function and some extensions are reported in (Cesa-Bianchi and Gentile, 2008; Kakade and Tewari, 2009). The main tool in this work is the use of martingale concentration inequalities (the Hoeffding-Azuma Inequality and the Friedman Inequality) to derive a bound on the average risk of the sequence of hypotheses generated by the learning algorithm in terms of a data-dependent statistic. Essentially, this relies on the fact that the differences V_t of the empirical loss $\ell(h_{t-1}, \boldsymbol{z}_t)$ and the true risk $\mathcal{R}(h_{t-1}) := \mathbb{E}_{\boldsymbol{z}}[\ell(h_{t-1}, \boldsymbol{z})]$ form a martingale sequence. Unfortunately, this property no longer holds for pairwise loss functions.

Of course, as mentioned for example in the work of Peel et al. (2010, Sec. 4.2), one can slightly adapt an existing online learning classification algorithm (e.g., perceptron), feeding it with data sequence $\breve{\boldsymbol{z}}_t := (\boldsymbol{z}_{2t-1}, \boldsymbol{z}_{2t})$ and modifying the update function accordingly. In this case, previous analysis (Cesa-Bianchi and Gentile, 2008) does apply. However, this does not make full use of the examples in the training sequence. In addition, empirical results show that this naive algorithm, which corresponds to the algorithm for online maximization of the area under the ROC curve (AUC) with a buffer size of one in (Zhao et al., 2011), is inferior to algorithms that retain some form of the history of the sequence. Alternatively, it is tempting to consider feeding the online algorithm with pairs $\breve{\boldsymbol{z}}_i^t = (\boldsymbol{z}_i, \boldsymbol{z}_t), i < t$ on each round. However, in this case, existing results would again fail because $\breve{\boldsymbol{z}}_i^t$ are not i.i.d. Hence, a natural question is whether we can prove data dependent generalization bounds based on the online pairwise loss.

This paper provides a positive answer to this question for a large family of pairwise loss functions. On each round t, we measure M_t, the average loss of h_{t-1} on examples $(\boldsymbol{z}_i, \boldsymbol{z}_t), i < t$. Let $\mathcal{M}^n$ denote the average loss, averaging M_t over $t \geqslant (1-c)n$ on a training sequence of length n where c is a small constant. The main result of this paper, provides a model selection mechanism to select one of the hypotheses of an *arbitrary* online learner, and states that the probability that the risk of the chosen hypothesis $\widehat{h}$ satisfies,

$$\mathcal{R}(\widehat{h}) \geqslant \mathcal{M}^n + \epsilon$$

is at most

$$2\left[\mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{64Lip(\phi)}\right)+1\right]\exp\left\{-\frac{(cn-1)\epsilon^2}{512}+2\ln n\right\}.$$

Here $\mathcal{N}(\mathcal{H},\eta)$ is the L_∞ covering number for the hypothesis class $\mathcal{H}$ and $Lip(\phi)$ is determined by the Lipschitz constant of the loss function (definitions and details are provided in the following sections). Our second main result is an analysis of a natural generalization of the perceptron algorithm to work with pairwise loss functions, that provides loss bounds in both the separable case and the inseparable case. As a byproduct, we also derive a new simple proof of the best L_1 based mistake bound for the perceptron algorithm in the inseparable case. Combining the two results we provide the first *online* algorithm with corresponding risk bound for bipartite ranking.

The rest of this paper is organized as follows. Section 2 defines the problem and states our main technical theorem and Section 3 provides a sketch of the proof. We provide model selection results and risk analysis for convex and general loss functions in Section 4. In Section 5, we describe our online algorithm for bipartite ranking and analyze it. Finally, we conclude the paper and discuss possible future directions in Section 6.

2. Main Technical Result

Given a sample $Z^n = \{\boldsymbol{z}_1,\cdots,\boldsymbol{z}_n\}$ where $\boldsymbol{z}_i$=$(\boldsymbol{x}_i, y_i)$ and a sequence of hypotheses $h_0,h_1,\cdots,$ h_n generated by an online learning algorithm, we define the sample statistic M^n as

$$\mathcal{M}^n(Z^n)=\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}M_t(Z^t), \qquad M_t(Z^t)=\frac{1}{t-1}\sum_{i=1}^{t-1}\ell\left(h_{t-1},\boldsymbol{z}_t,\boldsymbol{z}_i\right), \tag{2}$$

where $c_n=\lceil c\cdot n\rceil$ and $c\in(0,1)$ is a small positive constant. $M_t(Z^t)$ measures the performance of the hypothesis h_{t-1} on the next example $\boldsymbol{z}_t$ when paired with all previous examples. Note that instead of considering all the n generated hypotheses, we only consider the average of the hypotheses $h_{c_n-1},\cdots,h_{n-2}$ where the statistic M_t is reliable and the last two hypotheses h_{n-1},h_n are discarded for technical reasons. In the following, to simplify the notation, $\mathcal{M}^n$ denotes $\mathcal{M}^n(Z^n)$ and M_t denotes $M_t(Z^t)$.

As in (Cesa-Bianchi et al., 2004), our goal is to bound the average risk of the sequence of hypotheses in terms of $\mathcal{M}^n$, which can be obtained using the following theorem.

Theorem 1 *Assume the hypothesis space $(\mathcal{H},\|\cdot\|_\infty)$ is compact. Let $h_0,h_1,\cdots,h_n\in\mathcal{H}$ be the ensemble of hypotheses generated by an arbitrary online algorithm working with a pairwise loss function ℓ such that, $\ell(h,\boldsymbol{z}_1,\boldsymbol{z}_2)=\phi(y_1-y_2,h(\boldsymbol{x}_1)-h(\boldsymbol{x}_2))$, where $\phi:\mathbb{R}\times\mathbb{R}\to[0,1]$ is a Lipschitz function w.r.t. the second variable with a finite Lipschitz constant Lip(ϕ). Then, $\forall c>0,\forall\epsilon>0$, we have for sufficiently large n*

$$\mathbb{P}\left\{\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathcal{R}(h_{t-1})\geqslant\mathcal{M}^n+\epsilon\right\}\leqslant\left[2\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{32Lip(\phi)}\right)+1\right]\exp\left\{-\frac{(cn-1)\epsilon^2}{128}+\ln n\right\}. \tag{3}$$

Here the L_∞ covering number $\mathcal{N}(\mathcal{H},\eta)$ is defined to be the minimal ℓ in $\mathbb{N}$ such that there exist ℓ disks in $\mathcal{H}$ with radius η that cover $\mathcal{H}$. We make the following remarks.

Remark 2 *Let $\mathbb{E}_t$ denote $\mathbb{E}_{\boldsymbol{z}_t}[\cdot|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{t-1}]$. It can be seen that $\mathbb{E}_t[M_t] - \mathcal{R}(h_{t-1})$ is no longer a martingale sequence. Therefore, martingale concentration inequalities that are usually used in online-to-batch conversion do not directly yield the desired bound.*

Remark 3 *We need the assumption that the hypothesis space $\mathcal{H}$ is compact so that its covering number $\mathcal{N}(\mathcal{H}, \eta)$ is finite. As an example, suppose $\mathcal{X} \subset \mathbb{R}^d$ and the hypothesis space is the class of linear functions that lie within a ball $B_R(\mathbb{R}^d) = \{\boldsymbol{w} \in \mathbb{R}^d : \sup_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leqslant R\}$. It can be shown (see Cucker and Zhou, 2007, chap. 5) that the covering number is one if $\eta > R$ and otherwise*

$$\mathcal{N}(B_R, \eta) \leqslant \left(\frac{2R}{\eta} + 1 \right)^d . \tag{4}$$

Remark 4 *We say that $f(s,t)$ is Lipschitz w.r.t the second argument if $\forall s, |f(s,t_1) - f(s,t_2)| \leqslant Lip(f)\|t_1 - t_2\|$. This form of the pairwise loss function is not restrictive and is widely used. For example, in the supervised ranking problem, we can take the hinge loss as $\ell_{hinge}(h, \boldsymbol{z}_1, \boldsymbol{z}_2) = \phi(y_1 - y_2, h(\boldsymbol{x}_1) - h(\boldsymbol{x}_2)) = [1 - (h(\boldsymbol{x}_1) - h(\boldsymbol{x}_2))(y_1 - y_2)]_+$, which can be thought as a surrogate function for ℓ_{rank}. Since ϕ is not bounded, we define the bounded hinge loss using $\widetilde{\phi}(s,t) = \min([1 - st]_+, 1) \in [0,1]$ if $s \neq 0$ and 0 otherwise. We next show it is Lipschitz. This is trivial for $y = 0$. For $y \neq 0$, when the first argument is bounded a constant C, $\widetilde{\phi}(y, \cdot)$ satisfies $\left|\widetilde{\phi}(y, x_1) - \widetilde{\phi}(y, x_2)\right| \leqslant \left| [1 - yx_1]_+ - [1 - yx_2]_+ \right| \leqslant \|yx_1 - yx_2\| \leqslant C\|x_1 - x_2\|$. Alternatively, one can take the square loss, i.e. $\ell(h, \boldsymbol{z}_1, \boldsymbol{z}_2) = [1 - (h(\boldsymbol{x}_1) - h(\boldsymbol{x}_2))(y_1 - y_2)]^2$. If its support is bounded then ℓ is Lipschitz.*

3. Proof of the Main Technical Result

The proof is inspired by the work of (Cucker and Smale, 2002; Agarwal et al., 2005; Rudin, 2009). The proof makes use of the Hoeffding-Azuma inequality, McDiarmid's inequality, symmetrization techniques and covering numbers of compact spaces.

Proof [Proof of Theorem 1] By the definition of $\mathcal{M}^n$, we wish to bound

$$\mathbb{P}_{Z^n \sim \mathcal{D}^n} \left(\frac{1}{n - c_n} \sum_{t=c_n}^{n-1} \mathcal{R}(h_{t-1}) - \frac{1}{n - c_n} \sum_{t=c_n}^{n-1} M_t \geqslant \epsilon \right), \tag{5}$$

which can be rewritten as

$$\begin{aligned} &\mathbb{P} \left(\frac{1}{n - c_n} \sum_{t=c_n}^{n-1} \left[\mathcal{R}(h_{t-1}) - \mathbb{E}_t[M_t] \right] + \frac{1}{n - c_n} \sum_{t=c_n}^{n-1} \left[\mathbb{E}_t[M_t] - M_t \right] \geqslant \epsilon \right) \\ &\quad \leqslant \mathbb{P} \left(\frac{1}{n - c_n} \sum_{t=c_n}^{n-1} \left[\mathcal{R}(h_{t-1}) - \mathbb{E}_t[M_t] \right] \geqslant \frac{\epsilon}{2} \right) + \mathbb{P} \left(\frac{1}{n - c_n} \sum_{t=c_n}^{n-1} \left[\mathbb{E}_t[M_t] - M_t \right] \geqslant \frac{\epsilon}{2} \right). \end{aligned} \tag{6}$$

Thus, we need to bound two terms separately. The proof consists of four parts, as follows.

Step 1: Bounding the Martingale difference

First consider the second term in (6). We have that $V_t = (\mathbb{E}_t[M_t] - M_t)/(n - c_n)$ is a martingale difference sequence, i.e. $\mathbb{E}_t[V_t] = 0$. Since the loss function is bounded in $[0, 1]$, we have $|V_t| \leqslant 1/(n - c_n),\ t = 1, \cdots, n$. Therefore by the Hoeffding-Azuma inequality, $\sum_t V_t$ can be bounded such that

$$\mathbb{P}_{Z^n \sim \mathcal{D}^n}\left(\frac{1}{n - c_n}\sum_{t=c_n}^{n-1}\Big[\mathbb{E}_t[M_t] - M_t\Big] \geqslant \frac{\epsilon}{2}\right) \leqslant \exp\left\{-\frac{(1-c)n\epsilon^2}{2}\right\}. \tag{7}$$

Step 2: Symmetrization by a ghost sample Ξ^n

In this step we bound the first term in (6). Let us start with introducing a ghost sample $\Xi^n = \{\boldsymbol{\xi}_j\} = \{(\tilde{\boldsymbol{x}}_j, \tilde{y}_j)\}, j = 1, \cdots, n$ where each $\boldsymbol{\xi}_j$ follows the same distribution as $\boldsymbol{z}_j$. Recall the definition of M_t and define $\widetilde{M}_t$ as

$$M_t = \frac{1}{t-1}\sum_{j=1}^{t-1}\ell(h_{t-1}, \boldsymbol{z}_t, \boldsymbol{z}_j), \qquad \widetilde{M}_t = \frac{1}{t-1}\sum_{j=1}^{t-1}\ell(h_{t-1}, \boldsymbol{z}_t, \boldsymbol{\xi}_j). \tag{8}$$

The difference between $\widetilde{M}_t$ and M_t is that M_t is the sum of the loss incurred by h_{t-1} on the current instance $\boldsymbol{z}_t$ and all the previous examples $\boldsymbol{z}_j, j = 1, \cdots, t-1$ *on which h_{t-1} is trained*, while $\widetilde{M}_t$ is the loss incurred by the same hypothesis h_{t-1} on the current instance $\boldsymbol{z}_t$ and *an independent set of examples $\boldsymbol{\xi}_j, j = 1, \cdots, t-1$.*

Claim 1 *The following equation holds*

$$\mathbb{P}_{Z^n \sim \mathcal{D}^n}\left(\frac{1}{n - c_n}\sum_{t=c_n}^{n-1}[\mathcal{R}(h_{t-1}) - \mathbb{E}_t[M_t]] \geqslant \epsilon\right) \leqslant 2\mathbb{P}_{\substack{Z^n \sim \mathcal{D}^n \\ \Xi^n \sim \mathcal{D}^n}}\left(\frac{1}{n - c_n}\sum_{t=c_n}^{n-1}\Big[\mathbb{E}_t[\widetilde{M}_t] - \mathbb{E}_t[M_t]\Big] \geqslant \frac{\epsilon}{2}\right), \tag{9}$$

whenever $n/\log^2(n) \geqslant 2/(\epsilon(1-c))^2$.

Notice that the probability measure on the right hand side of (9) is on $Z^n \times \Xi^n$.

Proof [*Sketch of the proof of Claim 1*] It can be seen that the RHS (without the factor of 2) of (9) is at least

$$\begin{aligned}&\mathbb{P}_{\substack{Z^n \sim \mathcal{D}^n \\ \Xi^n \sim \mathcal{D}^n}}\left(\left\{\frac{1}{n - c_n}\sum_{t=c_n}^{n-1}[\mathcal{R}(h_{t-1}) - \mathbb{E}_t[M_t]] \geqslant \epsilon\right\}\bigcap\left\{\left|\frac{1}{n - c_n}\sum_{t=c_n}^{n-1}\Big[\mathbb{E}_t[\widetilde{M}_t] - \mathcal{R}(h_{t-1})\Big]\right| \leqslant \frac{\epsilon}{2}\right\}\right)\\ &= \mathbb{E}_{Z^n \sim \mathcal{D}^n}\left[\mathbb{I}_{\left\{\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}[\mathcal{R}(h_{t-1}) - \mathbb{E}_t[M_t]] \geqslant \epsilon\right\}} \cdot \mathbb{P}_{\Xi^n \sim \mathcal{D}^n}\left(\left|\frac{1}{n - c_n}\sum_{t=c_n}^{n-1}\Big[\mathbb{E}_t[\widetilde{M}_t] - \mathcal{R}(h_{t-1})\Big]\right| \leqslant \frac{\epsilon}{2}\middle| Z^n\right)\right].\end{aligned}$$

Since $\mathbb{E}_{\Xi^n \sim \mathcal{D}^n}\mathbb{E}_t[\widetilde{M}_t] = \mathcal{R}(h_{t-1})$, by Chebyshev's inequality

$$\mathbb{P}_{\Xi^n \sim \mathcal{D}^n}\left(\left|\frac{1}{n - c_n}\sum_{t=c_n}^{n-1}\Big[\mathbb{E}_t[\widetilde{M}_t] - \mathcal{R}(h_{t-1})\Big]\right| \leqslant \frac{\epsilon}{2}\middle| Z^n\right) \geqslant 1 - \frac{\mathbf{Var}\left\{\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathbb{E}_t[\widetilde{M}_t]\right\}}{\epsilon^2/4}. \tag{10}$$

To bound the variance, we first investigate the largest variation when changing one random variable $\boldsymbol{\xi}_j$ with others fixed. From (8), it can be easily seen that changing any of the $\boldsymbol{\xi}_j$ varies each $\mathbb{E}_t[\widetilde{M}_t]$, where $t > j$ by at most by $1/(t-1)$. Therefore, we can see that the variation of $\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathbb{E}_t[\widetilde{M}_t]$ regarding the jth example $\boldsymbol{\xi}_j$ is bounded by

$$c_j = \frac{1}{n-c_n}\left[\sum_{t=j+1}^{n}\frac{1}{t-1}\right] = \frac{1}{n-c_n}\left[\sum_{t=j}^{n-1}\frac{1}{t}\right] = \frac{1}{n-c_n}H_j(n).$$

The partial sum of the harmonic series $H_j(n) \leqslant \log(n),\ \forall j \geqslant 2$. Thus, by Theorem 9.3 in (Devroye et al., 1996), we have

$$\textbf{Var}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathbb{E}_t[\widetilde{M}_t]\right) \leqslant \frac{1}{4}\sum_{i=1}^{n}c_i^2 \leqslant \frac{1}{4(1-c)^2}\frac{\log^2(n)}{n}. \tag{11}$$

Thus, whenever $n/\log^2(n) \geqslant 2/(\epsilon(1-c))^2$, the LHS of (10) is greater or equal than $1/2$. This completes the proof of Claim 1. ■

Step 3: Uniform Convergence

In this step, we show how one can bound the RHS of (9) using uniform convergence techniques, McDiarmid's inequality and L_∞ covering number. Our task reduces to bound the following quantity

$$\mathbb{P}_{Z^n\sim\mathcal{D}^n,\Xi^n\sim\mathcal{D}^n}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_t[\widetilde{M}_t] - \mathbb{E}_t[M_t]\right] \geqslant \epsilon\right). \tag{12}$$

Here we want to bound the probability of the large deviation between the empirical performance of the ensemble of hypotheses on the sequence Z^n on which they are learned and on an independent sequence Ξ^n. Since h_t relies on $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t$ and is independent of $\{\boldsymbol{\xi}_t\}$, we resort to *uniform convergence techniques* to bound this probability. Define $L_t(h_{t-1}) = \mathbb{E}_t[\widetilde{M}_t] - \mathbb{E}_t[M_t]$. Thus we have

$$\begin{aligned}\mathbb{P}_{Z^n\sim\mathcal{D}^n,\Xi^n\sim\mathcal{D}^n}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}L_t(h_{t-1}) \geqslant \epsilon\right) &\leqslant \mathbb{P}\left(\sup_{\hat{h}_{c_n},\cdots,\hat{h}_{n-1}}\left[\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}L_t(\hat{h}_{t-1})\right] \geqslant \epsilon\right)\\ &\leqslant \sum_{t=c_n}^{n-1}\mathbb{P}_{Z^t\sim\mathcal{D}^t,\Xi^t\sim\mathcal{D}^t}\left(\sup_{\hat{h}\in\mathcal{H}}\left[L_t(\hat{h})\right] \geqslant \epsilon\right).\end{aligned} \tag{13}$$

To bound the RHS of (13), we start with the following lemma.

Lemma 5 *Given any function $f \in \mathcal{H}$ and any $t \geqslant 2$*

$$\mathbb{P}_{Z^t\sim\mathcal{D}^t,\Xi^t\sim\mathcal{D}^t}\left(L_t(f) \geqslant \epsilon\right) \leqslant \exp\left\{-\frac{(t-1)\epsilon^2}{2}\right\}. \tag{14}$$

The proof which is given in the appendix shows that $L_t(f)$ has a bounded variation of $1/(t-1)$ when changing each of its $2(t-1)$ variables and applies McDiarmid's inequality. Finally, our task is to bound $\mathbb{P}(\sup_{f \in \mathcal{H}} [L_t(f)] \geqslant \epsilon)$. Consider the simple case where the hypothesis space $\mathcal{H}$ is finite, then using the union bound, we immediately get the desired bound. Although $\mathcal{H}$ is not finite, a similar analysis goes through based on the assumption that $\mathcal{H}$ is compact. We will follow Cucker and Smale (2002) and show how this can be bounded. The next two lemmas (see proof of Lemma 6 in the appendix) are used to derive Lemma 8.

Lemma 6 *For any two functions $h_1, h_2 \in \mathcal{H}$, the following equation holds*

$$L_t(h_1) - L_t(h_2) \leqslant 4\mathit{Lip}(\phi)\|h_1 - h_2\|_\infty.$$

Lemma 7 *Let $\mathcal{H} = S_1 \cup \cdots \cup S_\ell$ and $\epsilon > 0$. Then*

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} L_t(h) \geqslant \epsilon\right) \leqslant \sum_{j=1}^{\ell} \mathbb{P}\left(\sup_{h \in S_j} L_t(h) \geqslant \epsilon\right)$$

Lemma 8 *For every $2 \leqslant t \leqslant n$, we have*

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} [L_t(h)] \geqslant \epsilon\right) \leqslant \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8\mathit{Lip}(\phi)}\right) \exp\left\{-\frac{(t-1)\epsilon^2}{8}\right\}. \tag{15}$$

Proof [*Proof of Lemma 8*] Let $\ell = \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{4\text{Lip}(\phi)}\right)$ and consider $h_1, \cdots, h_\ell$ such that the disks D_j centered at h_j and with radius $\frac{\epsilon}{4\text{Lip}(\phi)}$ cover $\mathcal{H}$. By Lemma 6, we have $|L_t(h) - L_t(h_j)| \leqslant 4\text{Lip}(\phi)\|h - h_j\|_\infty \leqslant \epsilon$. Thus, we get

$$\mathbb{P}\left(\sup_{h \in D_j} L_t(h) \geqslant 2\epsilon\right) \leqslant \mathbb{P}\left(L_t(h_j) \geqslant \epsilon\right)$$

Combining this with (14), and Lemma 7 and replacing ϵ by $\epsilon/2$, we have (15). ■

Combining (15) and (13), we have

$$\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1} L_t(h_{t-1}) \geqslant \epsilon\right) \leqslant \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8\text{Lip}(\phi)}\right) n \exp\left\{-\frac{(c_n-1)\epsilon^2}{8}\right\}. \tag{16}$$

This shows why we need to discard the first c_n hypotheses in the ensemble. If we include h_2 for example, according to (15), we have $\mathbb{P}(L_2(f) \geqslant \epsilon) \leqslant e^{-\epsilon^2/2}$. As n grows, this heavy term remains in the sum, and the desired bound cannot be obtained.

Step 4: Putting it all together

From (9) and (13) and substituting ϵ with $\epsilon/4$ in (16), we have

$$\mathbb{P}_{Z^n \sim \mathcal{D}^n}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}(\mathcal{R}(h_{t-1}) - \mathbb{E}_t[M_t]) \geqslant \frac{\epsilon}{2}\right) \leqslant 2\mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{32\text{Lip}(\phi)}\right) n \exp\left\{-\frac{(c_n-1)\epsilon^2}{128}\right\}. \tag{17}$$

From (17) and (7), accompanied with the fact that (17) decays faster than (7), we complete the proof for Theorem 1. ■

4. Model Selection

Following Cesa-Bianchi et al. (2004) our main tool for finding a good hypothesis from the ensemble of hypotheses generated by the online learner is to choose the one that has a small empirical risk. We measure the risk for h_t on the remaining $n-t$ examples, and penalize each h_t based on the number of examples on which it is evaluated, so that the resulting upper bound on the risk is reliable. Our construction and proofs (in the appendix) closely follow the ones in (Cesa-Bianchi et al., 2004), using large deviation results for U-statistics (see Clemençon et al., 2008, Appendix) instead of the Chernoff bound.

4.1. Risk Analysis for Convex losses

If the loss function ϕ is convex in its second argument and $\mathcal{Y}$ is convex, then we can use the average hypothesis $\bar{h} = \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} h_{t-1}$. It is easy to show that $\bar{h}$ achieves the desired bound, i.e.

$$\mathbb{P}\left(\mathcal{R}(\bar{h}) \geqslant M^n(Z^n) + \epsilon\right) \leqslant \left[2\mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{32\text{Lip}(\phi)}\right) + 1\right] \exp\left\{-\frac{(cn-1)\epsilon^2}{128} + \ln n\right\}.$$

4.2. Risk Analysis for General Losses

Define the empirical risk of hypothesis h_t on $\{\boldsymbol{z}_{t+1}, \cdots, \boldsymbol{z}_n\}$ as $\widehat{\mathcal{R}}(h_t, t+1)$

$$\widehat{\mathcal{R}}(h_t, t+1) = \binom{n-t}{2}^{-1} \sum_{k>i,\ i\geqslant t+1}^{n} \ell(h_t, \boldsymbol{z}_i, \boldsymbol{z}_k).$$

The hypothesis $\widehat{h}$ is chosen to minimize the following *penalized empirical risk*,

$$\widehat{h} = \operatorname*{argmin}_{c_n-1\leqslant t<n-1} (\widehat{\mathcal{R}}(h_t, t+1) + c_\delta(n-t)), \tag{18}$$

where

$$c_\delta(x) = \sqrt{\frac{1}{x-1}\ln\frac{2(n-c_n)(n-c_n+1)}{\delta}}.$$

Notice that we discard the last two hypotheses so that any $\widehat{\mathcal{R}}(h_t, t+1), c_n - 1 \leqslant t \leqslant n-2$ is well defined. The following theorem, which is the main result of this paper, shows that the risk of $\widehat{h}$ is bounded relative to $\mathcal{M}^n$.

Theorem 9 *Let $h_0, \cdots, h_n$ be the ensemble of hypotheses generated by an arbitrary online algorithm $\mathcal{A}$ working with a pairwise loss ℓ which satisfies the conditions given in Theorem 1. $\forall \epsilon > 0$, if the hypothesis is chosen via (18) with the confidence δ chosen as*

$$\delta = 2(n-c_n+1)\exp\left\{-\frac{(n-c_n)\epsilon^2}{128}\right\},$$

then, when n is sufficiently large, we have

$$\mathbb{P}\left(\mathcal{R}(\widehat{h}) \geqslant M^n + \epsilon\right) \leqslant 2\left[\mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{64Lip(\phi)}\right) + 1\right] \exp\left\{-\frac{(cn-1)\epsilon^2}{512} + 2\ln n\right\}.$$

5. Application: Online Algorithm for Bipartite Ranking

In the bipartite ranking problem we are given a sequence of labeled examples $\boldsymbol{z}_t = (\boldsymbol{x}_t, y_t) \in \mathbb{R}^d \times \{-1, +1\}, t = 1, \cdots, n$. Minimizing the *misranking loss* ℓ_{rank} under this setting is equivalent to maximizing the AUC, which measures the probability that f ranks a randomly drawn positive example higher than a randomly drawn negative example. This problem has been studied extensively in the batch setting, but the corresponding online problem has not been investigated until recently. Recently, Zhao et al. (2011) proposed an online algorithm using linear hypotheses for this problem based on reservoir sampling, and derived bounds on the expectation of the regret of this algorithm. Like previous work, Zhao et al. (2011) use the hinge loss (that bounds the 0-1 loss) to derive the regret bound. The hinge loss is Lipschitz, but it is not bounded and therefore not suitable for our risk bounds. Therefore, in the following we use a modified loss function where we bound the Hinge loss in $[0, 1]$ such that $\ell(f, \boldsymbol{z}_t, \boldsymbol{z}_j) = \widetilde{\phi}((y_t - y_j)/2, f(\boldsymbol{x}_t) - f(\boldsymbol{x}_j))$ where $\widetilde{\phi}$ is defined in Remark 4. Using this loss function together with Theorem 9 all we need is an online algorithm that minimizes $\mathcal{M}^n$ (or an upper bound of $\mathcal{M}^n$) and this guarantees generalization ability of the corresponding online learning algorithm. To this end, we propose the following perceptron-like algorithm, shown in Algorithm 5, and provide loss bounds for this algorithm. Notice that the algorithm does not treat each pair of examples separately, and instead for each $\boldsymbol{z}_t$ it makes a large combined update using its loss relative to all previous examples. Our algorithm corresponds to the algorithm of Zhao et al. (2011) with an infinite buffer, but it uses a different learning rate and different loss function which are important in our proofs.

```
Initialize: $\boldsymbol{w}_0 = \boldsymbol{0}$ repeat
    At the $t$-th iteration, receive a training instance $\boldsymbol{z}_t = (\boldsymbol{x}_t, y_t) \in \mathbb{R}^d \times \{-1, +1\}$.
    for $j \leftarrow 1$ to $t - 1$ do
        Calculate instantaneous loss $\ell_j^t = \ell(\boldsymbol{w}_{t-1}, \boldsymbol{z}_t, \boldsymbol{z}_j)$.
    end
    Update the weight vector such that
        $\boldsymbol{w}_t = \boldsymbol{w}_{t-1} + \frac{1}{t-1} \sum_{j=1}^{t-1} \ell_j^t y_t (\boldsymbol{x}_t - \boldsymbol{x}_j).$
until the last instance;
```

Algorithm 1: Online AUC Maximization (OAM) with Infinite Buffer.

Theorem 10 *Suppose we are given a sequence of examples $\boldsymbol{z}_t, t = 1, \cdots, n$, and let $\boldsymbol{u}$ be any unit vector. Assume $\max_t \|\boldsymbol{x}_t\| \leqslant R$ and define*

$$M = \sum_{t=2}^{n} \frac{1}{t-1} \left[\sum_{j=1}^{t-1} \ell_j^t \right], M^* = \sum_{t=2}^{n} \frac{1}{t-1} \left[\sum_{j=1}^{t-1} \hat{\ell}_j^t \right],$$

where $\hat{\ell}_j^t = \mathbb{I}_{y_t \neq y_j} \cdot \left[\gamma - \langle \boldsymbol{u}, \frac{1}{2}(y_t - y_j)(\boldsymbol{x}_t - \boldsymbol{x}_j) \rangle \right]_+$. That is, M^* is the cumulative average hinge loss $\boldsymbol{u}$ suffers on the sequence with margin γ. Then, after running Algorithm 5 on the sequence, we *have*

$$M \leqslant \left(\frac{\sqrt{4R^2+2} + \sqrt{\gamma M^*}}{\gamma} \right)^2 .$$

When the data is linearly separable by margin γ, (i.e. there exists an unit vector $\boldsymbol{u}$ such that $\hat{\ell}_j^t = 0, \forall t \leqslant n, j < t$), we have $M^* = 0$ and the bound is constant.

Proof [*Proof of Theorem 10*] First notice that $\boldsymbol{w}_0 = \boldsymbol{w}_1 = 0$ and we also have the following fact

$$\left[\gamma - \langle \boldsymbol{u}, \frac{1}{2}(y_t - y_j)(\boldsymbol{x}_t - \boldsymbol{x}_j) \rangle \right]_+ \geqslant \gamma - \langle \boldsymbol{u}, \frac{1}{2}(y_t - y_j)(\boldsymbol{x}_t - \boldsymbol{x}_j) \rangle,$$

which implies that when $y_t \neq y_j$, $\langle \boldsymbol{u}, y_t(\boldsymbol{x}_t - \boldsymbol{x}_j) \rangle \geqslant \gamma - \hat{\ell}_j^t$. On the other hand, when $y_t = y_j$, then $\hat{\ell}_j^t = 0$. Thus we can write

$$\begin{aligned}
\langle \boldsymbol{w}_t, \boldsymbol{u} \rangle &= \langle \boldsymbol{w}_{t-1}, \boldsymbol{u} \rangle + \frac{1}{t-1} \sum_{j=1}^{t-1} \ell_j^t \langle \boldsymbol{u}, y_t(\boldsymbol{x}_t - \boldsymbol{x}_j) \rangle \\
&\geqslant \langle \boldsymbol{w}_{t-1}, \boldsymbol{u} \rangle + \frac{1}{t-1} \sum_{j=1}^{t-1} \ell_j^t (\gamma - \hat{\ell}_j^t) = \langle \boldsymbol{w}_{t-1}, \boldsymbol{u} \rangle + \frac{\gamma}{t-1} \sum_{j=1}^{t-1} \ell_j^t - \frac{1}{t-1} \sum_{j=1}^{t-1} \ell_j^t \cdot \hat{\ell}_j^t \\
&\geqslant \langle \boldsymbol{w}_{t-1}, \boldsymbol{u} \rangle + \frac{\gamma}{t-1} \sum_{j=1}^{t-1} \ell_j^t - \frac{1}{t-1} \sum_{j=1}^{t-1} \hat{\ell}_j^t \qquad (\because \ell_j^t \in [0,1]) \\
\Rightarrow \quad \langle \boldsymbol{w}_t, \boldsymbol{u} \rangle &\geqslant \sum_{t=2}^{n} \left[\frac{\gamma}{t-1} \sum_{j=1}^{t-1} \ell_j^t - \frac{1}{t-1} \sum_{j=1}^{t-1} \hat{\ell}_j^t \right] = \gamma M - M^* . \qquad (19)
\end{aligned}$$

On the other hand, we have,

$$\begin{aligned}
\|\boldsymbol{w}_t\|^2 &= \|\boldsymbol{w}_{t-1}\|^2 + \frac{2}{t-1} \sum_{j=1}^{t-1} \ell_j^t \langle \boldsymbol{w}_{t-1}, y_t(\boldsymbol{x}_t - \boldsymbol{x}_j) \rangle + \left[\frac{1}{t-1} \sum_{j=1}^{t-1} \ell_j^t y_t (\boldsymbol{x}_t - \boldsymbol{x}_j) \right]^2 \\
&\leqslant \|\boldsymbol{w}_{t-1}\|^2 + \frac{2}{t-1} \sum_{j=1}^{t-1} \ell_j^t + 4R^2 \left(\frac{1}{t-1} \right)^2 \left(\sum_{j=1}^{t-1} \ell_j^t \right) \cdot \left(\sum_{j=1}^{t-1} \ell_j^t \right) \\
&\qquad (\because \ell_j^t > 0 \Rightarrow \langle \boldsymbol{w}_{t-1}, y_t(\boldsymbol{x}_t - \boldsymbol{x}_j) \rangle \leqslant 1) \\
&\leqslant \|\boldsymbol{w}_{t-1}\|^2 + \frac{2}{t-1} \sum_{j=1}^{t-1} \ell_j^t + 4R^2 \left(\frac{1}{t-1} \right)^2 \left(\sum_{j=1}^{t-1} \ell_j^t \right) \cdot (t-1) \qquad (\because \ell_j^t \in [0,1]) \\
&= \|\boldsymbol{w}_{t-1}\|^2 + (4R^2 + 2) \left[\frac{1}{t-1} \sum_{j=1}^{t-1} \ell_j^t \right] \\
\Rightarrow \quad \|\boldsymbol{w}_n\|^2 &\leqslant (4R^2+2) \sum_{t=2}^{n} \left[\frac{1}{t-1} \sum_{j=1}^{t-1} \ell_j^t \right] = (4R^2+2) M \qquad (20)
\end{aligned}$$

Combining (19) and (20), we have $(\gamma M - M^*)^2 \leqslant (4R^2+2)M$, which yields the desired result.
■

We therefore get the loss bound for the proposed algorithm.

Theorem 11 *Let $h_0, \cdots, h_{n-1}$ be the ensemble of hypotheses generated by Algorithm 5. $\forall \epsilon > 0$, if the hypothesis $\widehat{h}$ is chosen via (18) with the confidence δ chosen to be*

$$\delta = 2(n - c_n + 1) \exp \left\{ - \frac{(n - c_n)\epsilon^2}{128} \right\},$$

then the probability that

$$\mathcal{R}(\widehat{h}) \geqslant \frac{1}{n-c_n}\left[\left(\frac{\sqrt{4R^2+2}+\sqrt{\gamma M^*}}{\gamma}\right)^2\right]+\epsilon$$

is at most

$$2\left[\left(\frac{64R^2\sqrt{5n}}{\epsilon}+1\right)^d+1\right]\exp\left\{-\frac{(cn-1)\epsilon^2}{512}+2\ln n\right\}.$$

Proof [*Proof of Theorem 11*] By (20), we can easily see that $\|\boldsymbol{w}_t\| \leqslant \sqrt{n(4R^2+2)}, t=1,\cdots,n$, therefore we have $\|\boldsymbol{w}_t\|\cdot\|\boldsymbol{x}\| \leqslant R^2\sqrt{5n},\ \forall t \leqslant n$. Therefore, we can take the hypothesis space to be $\mathcal{H} = \{\boldsymbol{w} \in \mathbb{R}^d : \max_{\|\boldsymbol{x}\|\leqslant R} |\langle \boldsymbol{w}, \boldsymbol{x}\rangle| \leqslant R^2\sqrt{5n}\}$. By (4), the covering number can be calculated. On the other hand, from the definition in (2), it is easy to see that $\mathcal{M}^n \leqslant M/(n-c_n)$. Finally, combining Theorem 9 and Theorem 10 concludes the proof. ■

A natural criticism is that Algorithm 5 is not a real online algorithm due to the fact that the entire sample is stored and at each iteration t, the update requires $\mathcal{O}(t)$ time while online algorithms should have $\mathcal{O}(1)$ time per step. One can solve this problem by using the "reservoir sampling" techniques from (Zhao et al., 2011) (Random OAM). The idea is that at the t-th iteration, instead of keeping all previous $t-1$ examples, we keep a constant size buffer $\mathcal{B}_t$ that has a sample of the history. Zhao et al. (2011) gave a bound on the expectation of the cumulative loss $\mathcal{L} = \sum_t \sum_j \ell_j^t$. Translating their bound to our notation we get $\mathbb{E}[M] = M^* + \mathcal{O}(\sqrt{n})$ where the expectation is over randomly sampled instances in the buffer. Therefore, when the data are linearly separable, the cumulative loss given by this bound grows as $\mathcal{O}(\sqrt{n})$ which is worse than the bound we provided. In principle, one could turn the results of Zhao et al. (2011) into a high probability bound on M using the Chebyshev's inequality and then use Theorem 9 to analyze its risk. However, this does not yield exponential convergence as above. It would be interesting to investigate this further to improve the probabilistic analysis of the loss bound of Zhao et al. (2011), or integrate the buffer analysis into the risk bound of this paper to yield tighter results.

5.1. Mistake Bound for Perceptron

Interestingly, we can apply our proof strategy in Theorem 10 to analyze the Perceptron algorithm in the inseparable case. This yields the best known bound in terms of the one-norm of the hinge losses (given by (Gentile, 2003, Theorem 8) and (Shalev-Shwartz and Singer, 2005, Theorem 2)) using a simple direct proof.

Theorem 12 *(Gentile, 2003; Shalev-Shwartz and Singer, 2005) Let $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$ be a sequence of examples with $\|\boldsymbol{x}_i\| \leqslant R$. Let $\boldsymbol{u}$ be any unit vector and let $\gamma > 0$. Define the one-norm of the hinge losses as $D_1 = \sum_{t=1}^n \ell_t$, where $\ell_t = [\gamma - y_t\langle \boldsymbol{u}, \boldsymbol{x}_t\rangle]_+$. Then the number of mistakes the perceptron algorithm makes on this sequence is bounded by $\left(\frac{R+\sqrt{\gamma D_1}}{\gamma}\right)^2$.*

Proof Let $m_t = \mathbb{I}_{\mathrm{sgn}(\boldsymbol{w}_t\cdot\boldsymbol{x}_t)\neq y_t}$ so that the total number of mistakes is $M = \sum_t m_t$. Then, as usual, the upper bound is $\|\boldsymbol{w}_n\|^2 \leqslant R^2 M$. On the other hand, using the fact that $\ell_t = [\gamma - y_t\langle \boldsymbol{u}, \boldsymbol{x}_t\rangle]_+ \geqslant$

$\gamma - y_t\langle \boldsymbol{u}, \boldsymbol{x}_t\rangle$, which implies $y_t\langle \boldsymbol{u}, \boldsymbol{x}_t\rangle \geqslant \gamma - \ell_t$.. We therefore have the lower bound

$$\begin{aligned} \langle \boldsymbol{w}_{t+1}, \boldsymbol{u}\rangle &= \langle \boldsymbol{w}_t, \boldsymbol{u}\rangle + y_t\langle \boldsymbol{x}_t, \boldsymbol{u}\rangle m_t \geqslant \langle \boldsymbol{w}_t, \boldsymbol{u}\rangle + (\gamma - \ell_t)m_t \\ &= \langle \boldsymbol{w}_t, \boldsymbol{u}\rangle + \gamma m_t - \ell_t m_t \geqslant \langle \boldsymbol{w}_t, \boldsymbol{u}\rangle + \gamma m_t - \ell_t \qquad (\because m_t \leqslant 1) \\ \Rightarrow \quad \langle \boldsymbol{w}_n, \boldsymbol{u}\rangle &\geqslant \sum_{t=1}^{n} \gamma m_t - \sum_{t=1}^{n} \ell_t = \gamma M - D_1. \end{aligned} \tag{21}$$

Combing the upper bound R^2M with (21), we get $(\gamma M - D_1)^2 \leqslant R^2 M$. Solving the quadratic equation, we obtain the desired bound. ■

6. Conclusion and Future work

In this paper, we provide generalization bounds for online learners using pairwise loss functions and apply these to analyze the risk of an online Bipartite ranking algorithm. There are several directions for possible future work. From an empirical perspective, although the random Online AUC Maximization (OAM) is simple and easy to implement, it seems that it does not maintain buffers in an optimal way. Intuitively, one might want to store "support ranking vectors" that help to build the correct ranker instead of using a random buffer. We are currently exploring ideas on building a smart buffer to improve its performance.

From the theoretical point of view, one direction is to improve the current bounds to achieve faster convergence rates. Alternatively, one can analyze Algorithm 5 from a totally different point of view. Under the batch setting, Clemençon et al. (2008) already provide fast convergence rates for the *empirical risk minimizer*. Since Algorithm 5 is in fact a stochastic gradient descent algorithm to minimize the U-statistic, using online convex programming techniques (Zinkevich, 2003; Shalev-Shwartz, 2007), one can show that the regret is small. Combining this with the batch results automatically yields risk bounds for the algorithm. It is interesting to compare this approach to the one proposed in this paper in terms of the risk bounds that can be obtained. However, it is important to note that the approach in this paper is more general in two ways. First we only assume that the loss function is Lipschitz instead of convex. Second, the ensemble of hypotheses can be produced by an *arbitrary* online learning algorithm, not just stochastic gradient descent.

Acknowledgments

We would like to thank anonymous reviewers for their constructive comments. YW and DP also thank Nicolò Cesa-Bianchi for early discussion. YW and RK were partly supported by NSF grant IIS-0803409. Part of this research was done when YW was an intern at Akamai Technologies in 2011.

References

S. Agarwal and P. Niyogi. Stability and generalization of bipartite ranking algorithms. In *18th Annual Conference on Learning Theory*, 2005.

S. Agarwal and P. Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474, 2009.

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*, 6:393–425, 2005.

U. Brefeld and T. Scheffer. AUC maximizing support vector learning. In *ICML 2005 Workshop on ROC Analysis in Machine Learning*, 2005.

N. Cesa-Bianchi and C. Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, 54(1):386–390, 2008.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

S. Clemençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of u-statistics. *Annals of Statistics*, 36(2):844–874, 2008.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Am. Math. Soc.*, 39(1): 1–49, 2002.

F. Cucker and D. Zhou. *Learning theory: an approximation theory viewpoint*. Cambridge University Press, 2007.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer Verlag, 1996.

Y. Freund and R. Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37:277–296, 1999.

Y. Freund, R. Iyer, R. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

C. Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

T. Joachims. Optimizing search engines using clickthrough data. In *Eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

S. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. *Advances in Neural Information Processing Systems*, 2009.

M. Kearns, M. Li, L. Pitt, and L. G. Valiant. Recent results on boolean concept learning. In *Proceedings of the Fourth International Workshop on Machine Learning*, pages 337–352, 1987.

N. Littlestone. Mistake bounds and logarithmic linear-threshold learning algorithms. *PhD thesis, University of California at Santa Cruz*, 1990.

T. Peel, S. Anthoine, L. Ralaivola, et al. Empirical Bernstein inequalities for u-statistics. In *Neural Information Processing Systems (NIPS)*, 2010.

C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, 2009.

C. Rudin, C. Cortes, M. Mohri, and R. Schapire. Margin-based ranking meets boosting in the middle. *18th Annual Conference on Learning Theory*, 2005.

S. Shalev-Shwartz. Online learning: Theory, algorithms, and applications. *PhD thesis, Hebrew University*, 2007.

S. Shalev-Shwartz and Y. Singer. A new perspective on an old perceptron algorithm. *18th Anual Conference on Learning Theory*, 2005.

T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *18th Annual Conference on Learning Theory*, 2005.

P. Zhao, S. Hoi, R. Jin, and T. Yang. Online AUC Maximization. In *28th international conference on Machine learning*, 2011.

M. Zinkevich. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. In *20th international conference on Machine learning*, 2003.

Appendix A. Complete Proof of Claim 1

Proof [*Proof of Claim 1*] The required probability can be bounded as follows.

$$
\begin{aligned}
&\mathbb{P}_{Z^n\sim\mathcal{D}^n,\Xi^n\sim\mathcal{D}^n}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_t[\widetilde{M}_t]-\mathbb{E}_t[M_t]\right]\geqslant\frac{\epsilon}{2}\right)\\
&\geqslant\mathbb{P}_{Z^n\sim\mathcal{D}^n,\Xi^n\sim\mathcal{D}^n}\left(\left\{\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left(\mathcal{R}(h_{t-1})-\mathbb{E}_t[M_t]\right)\geqslant\epsilon\right\}\right.\\
&\qquad\qquad\left.\bigcap\left\{\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_t[\widetilde{M}_t]-\mathcal{R}(h_{t-1})\right]\right|\leqslant\frac{\epsilon}{2}\right\}\right)\\
&=\mathbb{E}_{Z^n\sim\mathcal{D}^n,\Xi^n\sim\mathcal{D}^n}\left[\mathbb{I}_{\left\{\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}(\mathcal{R}(h_{t-1})-\mathbb{E}_t[M_t])\geqslant\epsilon\right\}}\times\mathbb{I}_{\left\{\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_t[\widetilde{M}_t]-\mathcal{R}(h_{t-1})\right]\right|\leqslant\frac{\epsilon}{2}\right\}}\right]\\
&=\mathbb{E}_{Z^n\sim\mathcal{D}^n}\left[\mathbb{E}_{\Xi^n\sim\mathcal{D}^n}\left[\mathbb{I}_{\left\{\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}(\mathcal{R}(h_{t-1})-\mathbb{E}_t[M_t])\geqslant\epsilon\right\}}\times\mathbb{I}_{\left\{\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_t[\widetilde{M}_t]-\mathcal{R}(h_{t-1})\right]\right|\leqslant\frac{\epsilon}{2}\right\}}\middle|Z^n\right]\right]\\
&=\mathbb{E}_{Z^n\sim\mathcal{D}^n}\left[\mathbb{I}_{\left\{\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}(\mathcal{R}(h_{t-1})-\mathbb{E}_t[M_t])\geqslant\epsilon\right\}}\right.\\
&\qquad\qquad\left.\times\,\mathbb{P}_{\Xi^n\sim\mathcal{D}^n}\left(\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_t[\widetilde{M}_t]-\mathcal{R}(h_{t-1})\right]\right|\leqslant\frac{\epsilon}{2}\middle|Z^n\right)\right]
\end{aligned}
\tag{22}
$$

We next show that for sufficiently large n,

$$
\mathbb{P}_{\Xi^n\sim\mathcal{D}^n}\left(\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_t[\widetilde{M}_t]-\mathcal{R}(h_{t-1})\right]\right|\leqslant\frac{\epsilon}{2}\middle|Z^n\right)\geqslant\frac{1}{2},
$$

which combined with (22) implies (9). To begin with, we first show that the corresponding random variable has mean zero

$$
\begin{aligned}
&\mathbb{E}_{\Xi^n\sim\mathcal{D}^n}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_t[\widetilde{M}_t]-\mathcal{R}(h_{t-1})\right]\middle|Z^n\right)\\
&=\mathbb{E}_{\Xi^n\sim\mathcal{D}^n}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathbb{E}_t\left[\frac{1}{t-1}\sum_{j=1}^{t-1}\ell(h_{t-1},\boldsymbol{z}_t,\boldsymbol{\xi}_j)\right]-\mathcal{R}(h_{t-1})\middle|Z^n\right)\\
&=\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\frac{1}{t-1}\sum_{j=1}^{t-1}\mathbb{E}_{\boldsymbol{\xi}_j}\mathbb{E}_t[\ell(h_{t-1},\boldsymbol{z}_t,\boldsymbol{\xi}_j)|Z^t]-\mathcal{R}(h_{t-1})\right]\\
&=\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\left(\frac{1}{t-1}\sum_{j=1}^{t-1}\mathcal{R}(h_{t-1})\right)-\mathcal{R}(h_{t-1})\right]=0.
\end{aligned}
$$

Thus, we can use Chebyshev's inequality to bound the conditional probability as follows

$$
\mathbb{P}\left(\left\{\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_t[\widetilde{M}_t]-\mathcal{R}(h_{t-1})\right]\right|\leqslant\frac{\epsilon}{2}\right\}\middle|Z^n\right)\geqslant 1-\frac{\mathbf{Var}\left\{\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathbb{E}_t[\widetilde{M}_t]\right\}}{\epsilon^2/4}.
$$

To bound the variance, we resort to the following Theorem (see Devroye et al., 1996, Theorem 9.3)

Theorem 13 *Let $X_1, \cdots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_n, \boldsymbol{x}'} \left| f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_n) - f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}', \cdots, \boldsymbol{x}_n) \right| \leqslant c_i \quad \forall 1 \leqslant i \leqslant n.$$

Then

$$\textbf{\textit{Var}}(f(X_1, \cdots, X_n)) \leqslant \frac{1}{4} \sum_{i=1}^{n} c_i^2.$$

From (8), it can be easily seen that changing any of the $\boldsymbol{\xi}_j$ varies each $\mathbb{E}_t[\widetilde{M}_t]$, where $t > j$ by at most by $1/(t-1)$. Therefore, we can see that the variation of $\frac{1}{n-c_n} \sum_{t=c_n}^{n-1} \mathbb{E}_t[\widetilde{M}_t]$ regarding the jth example $\boldsymbol{\xi}_j$ is bounded by

$$c_j = \frac{1}{n - c_n} \left[\sum_{t=j+1}^{n} \frac{1}{t-1} \right] = \frac{1}{n - c_n} \left[\sum_{t=j}^{n-1} \frac{1}{t} \right] = \frac{1}{n - c_n} H_j(n).$$

The partial sum of the harmonic series $H_j(n) \leqslant \log(n)$, $\forall j > 2$. Thus, we have

$$\textbf{Var} \left(\frac{1}{n - c_n} \sum_{t=c_n}^{n-1} \mathbb{E}_t[\widetilde{M}_t] \right) \leqslant \frac{1}{4} \sum_{i=1}^{n} c_i^2 \leqslant \frac{1}{4(1-c)^2} \frac{\log^2(n)}{n}.$$

Thus, whenever $n / \log^2(n) \geqslant 2/(\epsilon(1-c))^2$, the LHS of (10) is greater or equal than $1/2$. This completes the proof of Claim 1. ■

Appendix B. Proof of Lemma 5

Proof [*Proof of Lemma 5*] To bound the probability, we use the McDiarmid's inequality.

Theorem 14 (McDiarmid's Inequality) *Let $X_1, \cdots, X_N$ be independent random variables with X_k taking values in a set A_k for each k. Let $\phi : (A_1 \times \cdots \times A_N) \to \mathbb{R}$ be such that*

$$\sup_{x_i \in A_i, x_k' \in A_k} |\phi(x_1, \cdots, x_N) - \phi(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_N)| \leqslant c_k.$$

Then for any $\epsilon > 0$,

$$\mathbb{P}\left\{ \phi(x_1, \cdots, x_N) - \mathbb{E}\phi(x_1, \cdots, x_N) \geqslant \epsilon \right\} \leqslant e^{-2\epsilon^2 / \sum_{k=1}^{N} c_k^2},$$

and

$$\mathbb{P}\left\{ |\phi(x_1, \cdots, x_N) - \mathbb{E}\phi(x_1, \cdots, x_N)| \geqslant \epsilon \right\} \leqslant 2e^{-2\epsilon^2 / \sum_{k=1}^{N} c_k^2}.$$

For any fixed $f \in \mathcal{H}$, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}_{1:t-1},\boldsymbol{\xi}_{1:t-1}}\left[L_t(f)\right] &= \frac{1}{t-1}\sum_{j=1}^{t-1}\mathbb{E}_{\boldsymbol{z}_{1:t-1},\boldsymbol{\xi}_{1:n-1}}\mathbb{E}_{\boldsymbol{z}}\left[\ell(f,\boldsymbol{z},\boldsymbol{\xi}_j)-\ell(f,\boldsymbol{z},\boldsymbol{z}_j)\right] \\ &= \frac{1}{t-1}\sum_{j=1}^{t-1}\left(\mathbb{E}_{\boldsymbol{\xi}_j}\mathbb{E}_{\boldsymbol{z}}[\ell(f,\boldsymbol{z},\boldsymbol{\xi}_j)]-\mathbb{E}_{\boldsymbol{z}_j}\mathbb{E}_{\boldsymbol{z}}[\ell(f,\boldsymbol{z},\boldsymbol{z}_j)]\right)=0\end{aligned}$$

Now, $L_t(f)$ is a function of $2(t-1)$ variables with each affecting its value at most by $c_i = 1/(t-1), i = 1, 2, \cdots, 2(t-1)$. Thus, we have $\sum_{i=1}^{2(t-1)} c_i^2 = \frac{1}{t-1}$. Finally, using the McDiarmid's inequality, we get

$$\mathbb{P}_{Z^t\sim\mathcal{D}^t,\Xi^t\sim\mathcal{D}^t}\left(L_t(f)\geqslant\epsilon\right)\leqslant\exp\left\{-\frac{(t-1)\epsilon^2}{2}\right\}.$$

■

Appendix C. Proof of Lemma 6

Proof [*Proof of Lemma 6*] From the definition of L_t and the assumption on ϕ we have

$$\begin{aligned}L_t(h_1)-L_t(h_2) &= \frac{1}{t-1}\sum_{j=1}^{t-1}\Big[\mathbb{E}_{\boldsymbol{z}}[\ell(h_1,\boldsymbol{z},\boldsymbol{\xi}_j)-\ell(h_1,\boldsymbol{z},\boldsymbol{z}_j)]-\mathbb{E}_{\boldsymbol{z}}[\ell(h_2,\boldsymbol{z},\xi_j)-\ell(h_2,\boldsymbol{z},\boldsymbol{z}_j)]\Big] \\ &= \frac{1}{t-1}\sum_{j=1}^{t-1}\mathbb{E}_{\boldsymbol{z}}\Big\{\big[\phi(y-\tilde{y}_j,h_1(\boldsymbol{x})-h_1(\tilde{\boldsymbol{x}}_j))-\phi(y-y_j,h_1(\boldsymbol{x})-h_1(\boldsymbol{x}_j))\big] \\ &\qquad -\big[\phi(y-\tilde{y}_j,h_2(\boldsymbol{x})-h_2(\tilde{\boldsymbol{x}}_j))-\phi(y-y_j,h_2(\boldsymbol{x})-h_2(\boldsymbol{x}_j))\big]\Big\} \\ &= \frac{1}{t-1}\sum_{j=1}^{t-1}\mathbb{E}_{\boldsymbol{z}}\Big\{\big[\phi(y-\tilde{y}_j,h_1(\boldsymbol{x})-h_1(\tilde{\boldsymbol{x}}_j)-\phi(y-\tilde{y}_j,h_2(\boldsymbol{x})-h_2(\tilde{\boldsymbol{x}}_j)\big] \qquad . \\ &\qquad -\big[\phi(y-y_j,h_1(\boldsymbol{x})-h_1(\boldsymbol{x}_j))-\phi(y-y_j,h_2(\boldsymbol{x})-h_2(\boldsymbol{x}_j))\big]\Big\} \\ &\leqslant \frac{1}{t-1}\sum_{j=1}^{t-1}\mathbb{E}_{\boldsymbol{z}}\Big\{\text{Lip}(\phi)\Big|\Big[h_1(\boldsymbol{x})-h_1(\tilde{\boldsymbol{x}}_j)\Big]-\Big[h_2(\boldsymbol{x})-h_2(\tilde{\boldsymbol{x}}_j)\Big]\Big| \\ &\qquad +\text{Lip}(\phi)\Big|\Big[h_1(\boldsymbol{x})-h_1(\boldsymbol{x}_j)\Big]-\Big[h_2(\boldsymbol{x})-h_2(\boldsymbol{x}_j)\Big]\Big|\Big\} \\ &\leqslant \frac{1}{t-1}\sum_{j=1}^{t-1}\Big[4\text{Lip}(\phi)\sup_{\boldsymbol{x}'}\big|h_1(\boldsymbol{x}')-h_2(\boldsymbol{x}')\big|\Big]=4\text{Lip}(\phi)\|h_1-h_2\|_\infty\end{aligned}$$

■

Appendix D. Proof for the Risk Bound of Convex Losses in Section 4.1

Proof Using Jensen's inequality, we have

$$\begin{aligned}\mathcal{R}(\bar{h}) &= \mathbb{E}_{\boldsymbol{z}}\mathbb{E}_{\boldsymbol{z}'}\left[\phi\left(y-y', \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} h_{t-1}(\boldsymbol{x}) - \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} h_{t-1}(\boldsymbol{x}')\right)\right] \\ &= \mathbb{E}_{\boldsymbol{z}}\mathbb{E}_{\boldsymbol{z}'}\left[\phi\left(y-y', \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} \left[h_{t-1}(\boldsymbol{x}) - h_{t-1}(\boldsymbol{x}')\right]\right)\right] \\ &\leqslant \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} \mathbb{E}_{\boldsymbol{z}}\mathbb{E}_{\boldsymbol{z}'}[\phi(y-y', h_{t-1}(\boldsymbol{x}) - h_{t-1}(\boldsymbol{x}'))] \\ &= \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} \mathbb{E}_{\boldsymbol{z}}\mathbb{E}_{\boldsymbol{z}'}[\ell(h_{t-1}, \boldsymbol{z}, \boldsymbol{z}')] = \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} \mathcal{R}(h_{t-1}).\end{aligned}$$

Combining with Theorem 1, we have

$$\mathbb{P}\left(\mathcal{R}(\bar{h}) \geqslant M^n(Z^n) + \epsilon\right) \leqslant \left[2\mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{32\text{Lip}(\phi)}\right) + 1\right] \exp\left\{-\frac{(cn-1)\epsilon^2}{128} + \ln n\right\}.$$

■

Appendix E. Proof of Theorem 9

Proof [*Proof of Theorem 9*] The proof is adapted from the proof for Theorem 4 in (Cesa-Bianchi et al., 2004). The main difference is that instead of using the Chernoff bound we use a large deviation bound for the U-statistic as follows.

Lemma 15 *(see Clemençon et al., 2008, Appendix) Suppose we have i.i.d. random variables $X_1, \cdots, X_n \in \mathcal{X}$ and the $U-$statistic is defined as*

$$U_n = \frac{1}{n(n-1)}\sum_{i\neq j}^{n} q(X_i, X_j) = \frac{2}{n(n-1)}\sum_{i>j}^{n} q(X_i, X_j),$$

where the kernel $q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric real-valued function. Then we have,

$$\mathbb{P}(|U_n - \mathbb{E}[U_n]| \geqslant \epsilon) \leqslant 2\exp\{-(n-1)\epsilon^2\}. \tag{23}$$

Therefore, by (23), we have

$$\mathbb{P}\left(|\widehat{\mathcal{R}}(h_t, t+1) - \mathcal{R}(h_t)| \geqslant \epsilon\right) \leqslant 2\exp\{-(n-t-1)\epsilon^2\},$$

or equivalently,

$$\mathbb{P}\left(\left|\widehat{\mathcal{R}}(h_t, t+1) - \mathcal{R}(h_t)\right| \geqslant \sqrt{\frac{1}{n-t-1}\ln\frac{2}{\delta}}\right) \leqslant \delta. \tag{24}$$

By the definition of c_δ and (24), one can see that

$$\mathbb{P}(\left|\widehat{\mathcal{R}}(h_t, t+1) - \mathcal{R}(h_t)\right| > c_\delta(n-t)) \leqslant \frac{\delta}{(n-c_n)(n-c_n+1)}. \tag{25}$$

Next, we show the following lemma,

Lemma 16 *Let $h_0, \cdots, h_{n-1}$ be the ensemble of hypotheses generated by an arbitrary online algorithm $\mathcal{A}$ working with a pairwise loss ℓ which satisfies the conditions given in Theorem 1. Then for any $0 < \delta \leqslant 1$, we have*

$$\mathbb{P}\left(\mathcal{R}(\widehat{h}) \geqslant \min_{c_n-1 \leqslant t < n-1} \left(\mathcal{R}(h_t) + 2c_\delta(n-t)\right)\right) \leqslant \delta. \tag{26}$$

Proof [*Proof of Lemma 16*] The proof closely follows the proof of Lemma 3 in (Cesa-Bianchi et al., 2004) and is given for the sake of completeness. Let

$$T^* = \operatorname*{argmin}_{c_n-1 \leqslant t < n-1} \left(\mathcal{R}(h_t) + 2c_\delta(n-t)\right),$$

and $h^* = h_{T^*}$ is the corresponding hypothesis that minimizes the penalized true risk and let $\widehat{\mathcal{R}}^*$ to be the penalized empirical risk of h_{T^*}, i.e.

$$\widehat{\mathcal{R}}^* = \widehat{\mathcal{R}}(h_{T^*}, T^*+1).$$

Set, for brevity

$$\widehat{\mathcal{R}}_t = \widehat{\mathcal{R}}(h_t, t+1),$$

and let

$$\widehat{T} = \operatorname*{argmin}_{c_n-1 \leqslant t < n-1} (\widehat{\mathcal{R}}_t + c_\delta(n-t)),$$

where $\widehat{h}$ defined in (18) coincides with $h_{\widehat{T}}$. With this notation, and since

$$\widehat{\mathcal{R}}_{\widehat{T}} + c_\delta(n-\widehat{T}) \leqslant \widehat{\mathcal{R}}^* + c_\delta(n-T^*)$$

holds with certainty, we can write

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{R}(\widehat{h}) > \mathcal{R}(h^*) + \mathcal{E}\right) \\
&\quad = \mathbb{P}\left(\mathcal{R}(\widehat{h}) > \mathcal{R}(h^*) + \mathcal{E},\ \widehat{\mathcal{R}}_{\widehat{T}} + c_\delta(n-\widehat{T}) \leqslant \widehat{\mathcal{R}}^* + c_\delta(n-T^*)\right) \\
&\quad \leqslant \sum_{t=c_n-1}^{n-2} \mathbb{P}\left(\mathcal{R}(h_t) > \mathcal{R}(h^*) + \mathcal{E},\ \widehat{\mathcal{R}}_t + c_\delta(n-t) \leqslant \widehat{\mathcal{R}}^* + c_\delta(n-T^*)\right)
\end{aligned}$$

where $\mathcal{E}$ is a positive-valued random variable to be specified. Now if

$$\widehat{\mathcal{R}}_t + c_\delta(n-t) \leqslant \widehat{\mathcal{R}}^* + c_\delta(n-T^*)$$

holds, then at least one of the following three conditions:

$$\begin{aligned}
&\widehat{\mathcal{R}}_t \leqslant \mathcal{R}(h_t) - c_\delta(n-t)\\
&\widehat{\mathcal{R}}^* > \mathcal{R}(h^*) + c_\delta(n-T^*)\\
&\mathcal{R}(h_t) - \mathcal{R}(h^*) < 2c_\delta(n-T^*)
\end{aligned}$$

must hold. Therefore, for any fixed t, we can write

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{R}(h_t) > \mathcal{R}(h^*) + \mathcal{E},\ \widehat{\mathcal{R}}_t + c_\delta(n-t) \leqslant \widehat{\mathcal{R}}^* + c_\delta(n-T^*)\right)\\
&\quad\leqslant \mathbb{P}\left(\widehat{\mathcal{R}}_t \leqslant \mathcal{R}(h_t) - c_\delta(n-t)\right) + \mathbb{P}\left(\widehat{\mathcal{R}}^* > \mathcal{R}(h^*) + c_\delta(n-T^*)\right)\\
&\quad\quad + \mathbb{P}\left(\mathcal{R}(h_t) - \mathcal{R}(h^*) < 2c_\delta(n-T^*), \mathcal{R}(h_t) > \mathcal{R}(h^*) + \mathcal{E}\right).
\end{aligned}$$

The last term is zero if we choose $\mathcal{E} = 2c_\delta(n-T^*)$. Hence, we can write

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{R}(\widehat{h}) > \mathcal{R}(h^*) + 2c_\delta(n-T^*)\right)\\
&\quad\leqslant \sum_{t=c_n-1}^{n-2} \mathbb{P}\left(\widehat{\mathcal{R}}_t \leqslant \mathcal{R}(h_t) - c_\delta(n-t)\right) + (n-c_n)\mathbb{P}\left(\widehat{\mathcal{R}}^* > \mathcal{R}(h^*) + c_\delta(n-T^*)\right)\\
&\quad\leqslant (n-c_n) \times \frac{\delta}{(n-c_n)(n-c_n+1)} \qquad \text{(By (25).)}\\
&\quad\quad + (n-c_n)\left[\sum_{t=c_n-1}^{n-2} \mathbb{P}\left(\widehat{\mathcal{R}}_t > \mathcal{R}(h_t) + c_\delta(n-t)\right)\right]\\
&\quad\leqslant \frac{\delta}{n-c_n+1} + (n-c_n)^2 \times \frac{\delta}{(n-c_n)(n-c_n+1)} \qquad \text{(By (25).)}\\
&\quad= \frac{\delta}{n-c_n+1} + \frac{(n-c_n)\delta}{n-c_n+1} = \delta.
\end{aligned}$$

■

Therefore, we know that

$$\mathbb{P}\left(\mathcal{R}(\widehat{h}) \geqslant \min_{c_n-1\leqslant t<n-1} \left(\mathcal{R}(h_t) + 2c_\delta(n-t)\right)\right) \leqslant \delta. \tag{27}$$

The next step is to show that with high probability $\min_{c_n-1\leqslant t<n-1}(\mathcal{R}(h_t)+2c_\delta(n-t))$ is close to M^n. To begin with, notice that

$$\begin{aligned}
&\min_{c_n-1\leqslant t<n-1}(\mathcal{R}(h_t)+2c_\delta(n-t))\\
&\quad=\min_{c_n-1\leqslant t<n-1}\min_{t\leqslant i<n-1}(\mathcal{R}(h_i)+2c_\delta(n-i))\\
&\quad\leqslant\min_{c_n-1\leqslant t<n-1}\frac{1}{n-1-t}\sum_{i=t}^{n-2}(\mathcal{R}(h_i)+2c_\delta(n-i))\\
&\quad=\min_{c_n-1\leqslant t<n-1}\left(\frac{1}{n-1-t}\sum_{i=t}^{n-2}\mathcal{R}(h_i)\right.\\
&\qquad\qquad\left.+\frac{2}{n-1-t}\sum_{i=t}^{n-2}\sqrt{\frac{1}{n-i-1}\ln\frac{2(n-c_n)(n-c_n+1)}{\delta}}\right)\\
&\quad\leqslant\min_{c_n-1\leqslant t<n-1}\left(\frac{1}{n-1-t}\sum_{i=t}^{n-2}\mathcal{R}(h_i)+\frac{2}{n-1-t}\sum_{i=t}^{n-2}\sqrt{\frac{2}{n-i-1}\ln\frac{2(n-c_n+1)}{\delta}}\right)\\
&\quad\leqslant\min_{c_n-1\leqslant t<n-1}\left(\frac{1}{n-1-t}\sum_{i=t}^{n-2}\mathcal{R}(h_i)+4\sqrt{\frac{2}{n-t-1}\ln\frac{2(n-c_n+1)}{\delta}}\right)
\end{aligned}$$

where the last equality holds because $\sum_{i=1}^{n-t-1}\sqrt{1/i}\leqslant 2\sqrt{n-t-1}$ (see Cesa-Bianchi et al., 2004, Sec. 2.B). Define

$$M_{m,n}=\frac{1}{n-m}\sum_{t=m}^{n-1}M_t(Z^t).$$

From Theorem 1, one can see that for each $t=c_n-1,\cdots,n-2$,

$$\mathbb{P}\left(\frac{1}{n-1-t}\sum_{i=t}^{n-2}\mathcal{R}(h_i)\geqslant M_{t,n}+\epsilon\right)\leqslant\left[2\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{32\text{Lip}(\phi)}\right)+1\right]\exp\left\{-\frac{(t-1)\epsilon^2}{128}+\ln n\right\}. \tag{28}$$

Then, set for brevity,

$$K_t=M_{t,n}+4\sqrt{\frac{2}{n-t-1}\ln\frac{2(n-c_n+1)}{\delta}}+\epsilon.$$

Using the fact that if $\min(a_1,a_2)\leqslant\min(b_1,b_2)$ then either $a_1\leqslant b_1$ or $a_2\leqslant b_2$, we can write

$$
\begin{aligned}
&\mathbb{P}\left(\min_{c_n-1\leqslant t<n-1}\left(\mathcal{R}(h_t)+2c_\delta(n-t)\right)\geqslant\min_{c_n-1\leqslant t<n-1}K_t\right)\\
&\quad\leqslant\mathbb{P}\Bigg(\min_{c_n-1\leqslant t<n-1}\Bigg(\frac{1}{n-1-t}\sum_{i=t}^{n-2}\mathcal{R}(h_i)\\
&\qquad\qquad+4\sqrt{\frac{2}{n-t-1}\ln\frac{2(n-c_n+1)}{\delta}}\Bigg)\geqslant\min_{c_n-1\leqslant t<n-1}K_t\Bigg)\\
&\quad\leqslant\sum_{t=c_n-1}^{n-2}\mathbb{P}\left(\frac{1}{n-1-t}\sum_{i=t}^{n-2}\mathcal{R}(h_i)+4\sqrt{\frac{2}{n-t-1}\ln\frac{2(n-c_n+1)}{\delta}}\geqslant K_t\right)\\
&\quad=\sum_{t=c_n-1}^{n-2}\mathbb{P}\left(\frac{1}{n-1-t}\sum_{i=t}^{n-2}\mathcal{R}(h_i)\geqslant M_{t,n}+\epsilon\right)\\
&\quad\leqslant(n-c_n-1)\left[2\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{32\text{Lip}(\phi)}\right)\right]\exp\left\{-\frac{(cn-1)\epsilon^2}{128}+\ln n\right\}\qquad\text{(By (28).)}\\
&\quad\leqslant\left[2\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{32\text{Lip}(\phi)}\right)\right]\exp\left\{-\frac{(cn-1)\epsilon^2}{128}+2\ln n\right\}.
\end{aligned}
$$

Therefore, using (27), we get

$$
\begin{aligned}
&\mathbb{P}\left(\mathcal{R}(\widehat{h})\geqslant\min_{c_n-1\leqslant t<n-1}\left(M_{t,n}+4\sqrt{\frac{2}{n-t-1}\ln\frac{2(n-c_n+1)}{\delta}}\right)+\epsilon\right)\\
&\qquad\leqslant\delta+\left[2\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{32\text{Lip}(\phi)}\right)\right]\exp\left\{-\frac{(cn-1)\epsilon^2}{128}+2\ln n\right\},
\end{aligned}
$$

which, in particular, leads to

$$
\begin{aligned}
&\mathbb{P}\left(\mathcal{R}(\widehat{h})\geqslant M^n+4\sqrt{\frac{2}{n-c_n}\ln\frac{2(n-c_n+1)}{\delta}}+\epsilon\right)\\
&\qquad\leqslant\delta+\left[2\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{32\text{Lip}(\phi)}\right)\right]\exp\left\{-\frac{(cn-1)\epsilon^2}{128}+2\ln n\right\}.
\end{aligned}
$$

By substituting ϵ with $\epsilon/2$ and choosing δ as in the statement of Theorem 1, that is, satisfying $4\sqrt{\frac{2}{n-c_n}\ln\frac{2(n-c_n+1)}{\delta}}=\frac{\epsilon}{2}$, we have for any $c>0$,

$$
\begin{aligned}
&\mathbb{P}\left(\mathcal{R}(\widehat{h})\geqslant M^n+\epsilon\right)\leqslant 2(n-c_n+1)\exp\left\{-\frac{(n-c_n)\epsilon^2}{128}\right\}+\left[2\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{64\text{Lip}(\phi)}\right)+1\right]\times\\
&\exp\left\{-\frac{(cn-1)\epsilon^2}{512}+2\ln n\right\}\leqslant 2\left[\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{64\text{Lip}(\phi)}\right)+1\right]\exp\left\{-\frac{(cn-1)\epsilon^2}{512}+2\ln n\right\}
\end{aligned}
$$

■