Generalization Bounds for Regularized Pairwise Learning

Yunwen Lei[1], **Shao-Bo Lin**[2], **Ke Tang**[1]
[1] Shenzhen Key Laboratory of Computational Intelligence, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen
[2] Department of Mathematics, Wenzhou University, Wenzhou
leiyw@sustc.edu.cn, sblin1983@gmail.com, tangk3@sustc.edu.cn

Abstract

Pairwise learning refers to learning tasks with the associated loss functions depending on pairs of examples. Recently, pairwise learning has received increasing attention since it covers many machine learning schemes, e.g., metric learning, ranking and AUC maximization, in a unified framework. In this paper, we establish a unified generalization error bound for regularized pairwise learning without either Bernstein conditions or capacity assumptions. We apply this general result to typical learning tasks including distance metric learning and ranking, for each of which our discussion is able to improve the state-of-the-art results.

1 Introduction

Recently, there is a growing interest in studying a large family of machine learning problems called pairwise learning problems. Unlike traditional learning problems whose loss functions depend only on a single example (e.g., classification and regression), pairwise learning refers to learning tasks for which the associated loss function involves a pair of examples. Specifically, for any two examples $(x, y), (\tilde{x}, \tilde{y}) \in \mathbb{R}^d \times \mathbb{R}$, the loss function for pairwise learning often takes the form $V(h, (x, y), (\tilde{x}, \tilde{y}))$ for a hypothesis function $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Many learning tasks can be cast into the framework of pairwise learning, including ranking [Rejchel, 2012; Kriukova *et al.*, 2016], metric learning [Xing *et al.*, 2003; Cao *et al.*, 2016], AUC maximization [Cortes and Mohri, 2004; Gao and Zhou, 2015; Gao *et al.*, 2013; Zhao *et al.*, 2011], gradient learning [Mukherjee and Zhou, 2006] and learning under minimum error entropy criterion [Hu *et al.*, 2015], etc. For example, supervised metric learning aims to find a Mahalanobis metric $d_{\mathbf{w}}(x, \tilde{x}) = (x - \tilde{x})^\top \mathbf{w}(x - \tilde{x})$ encoded by a semi-positive matrix $\mathbf{w} \in \mathbb{S}^{d\times d}$ to bring examples with similar labels together while keeping examples with different labels apart [Xing *et al.*, 2003], where $\mathbb{S}^{d\times d}$ is the class of all positive semi-definite matrices in $\mathbb{R}^{d\times d}$. In this case, a common loss function $V(d_{\mathbf{w}}, (x, y), (\tilde{x}, \tilde{y})) = g(y\tilde{y}(1 - d_{\mathbf{w}}(x, \tilde{x})))$ involves two examples $(x, y), (\tilde{x}, \tilde{y})$, where $g : \mathbb{R} \to \mathbb{R}_+$ is convex for which a typical choice is the hinge loss $g(t) = \max(1 - t, 0)$ [Cao *et al.*, 2016].

Motivated by growing interests in pairwise learning, generalization analysis of different pairwise learning machines has been conducted to better understand their practical behavior [Kar *et al.*, 2013; Ying and Zhou, 2016; Christmann and Zhou, 2016; Clémençon *et al.*, 2008; Rejchel, 2012]. A difficulty in learning theory analysis of pairwise learning consists in the fact that the empirical error can not be written as a summation of independent and identically distributed (i.i.d.) random variables, rendering standard techniques in the i.i.d. case not applicable in this context [Sridharan *et al.*, 2009; Bartlett *et al.*, 2005]. For these learning problems with coupled examples, existing studies use either techniques in U-process to derive uniform convergence bounds [Rejchel, 2012; Clémençon *et al.*, 2008; Cao *et al.*, 2016; Zhao *et al.*, 2017; Lei and Ying, 2016] or algorithm stability/robustness to establish algorithm-specific bounds [Bellet and Habrard, 2015; Jin *et al.*, 2009; Agarwal and Niyogi, 2009]. However, existing studies on pairwise learning problems are not quite satisfactory in the following three aspects. Firstly, these generalization bounds are mostly derived for different specific instantiations of pairwise learning problems, and a unified framework to study generalization errors for regularized pairwise learning is still lacking. Secondly, most of these discussions only consider estimation errors. Thirdly, these estimation error bounds either are slow [Jin *et al.*, 2009; Bellet and Habrard, 2015; Agarwal and Niyogi, 2009; Cao *et al.*, 2016] or require capacity assumptions on hypothesis spaces and Bernstein conditions for an application of Talagrand's inequality [Rejchel, 2012; Clémençon *et al.*, 2008].

In this paper, we provide a unified analysis for regularized pairwise learning by showing that the regularized generalization error of the estimator would converge to the optimal value at a rate of the order $O(1/n)$, where n is the number of training examples. Our discussion requires neither Bernstein conditions on the bias and variance nor capacity assumptions on the hypothesis spaces. The property that the empirical error is a U-statistic makes the argument in the i.i.d. case not applicable to our context, and we bypass this obstacle by resorting to established techniques in the U-process. Based on this, we develop generalization error bounds for regularized pairwise learning which also take approximation errors into consideration. We apply our general results to metric learning and ranking, for each of which our general analysis can imply generalization bounds tighter than the state of the art.

2 Related Work

Here, we review related work on generalization analysis of pairwise learning algorithms based on different approaches.

Generalization error bounds were established for regularized metric learning [Jin *et al.*, 2009] and ranking [Agarwal and Niyogi, 2009] based on algorithmic stability. The basic idea is to use strong convexity of the objective function in regularized pairwise learning problems to show that the learned model would change slightly if a single training example is replaced by another one. Based on this, McDiarmid's inequality is applied to establish generalization bounds. However, this approach can only yield a suboptimal estimation bound $O(1/(\lambda\sqrt{n}))$, where λ is the regularization parameter.

The tool of U-process was also used in generalization analysis of regularized metric learning [Cao *et al.*, 2016] and ranking [Clémençon *et al.*, 2008; Rejchel, 2012]. The basic idea is to use symmetry of U-statistics to control supremum of a U-process in generalization analysis by the supremum of a Rademacher process, the latter of which can be bounded by standard techniques in the i.i.d. setting. However, existing studies with this approach either imply a suboptimal estimation bound $O(1/(\sqrt{\lambda n}))$ or require both Bernstein conditions and capacity assumptions for a fast learning rate.

Integral operator was used to establish learning rates for regularized least squares ranking [Zhao *et al.*, 2017]. The basic idea is to show that the involved optimization problem has a closed-form solution in terms of integral operators, which, however, applies only to the least squares loss.

Recently, regret bounds for online pairwise learning algorithms were established based on online-to-batch conversion together with covering numbers [Wang *et al.*, 2012] and Rademacher complexities [Kar *et al.*, 2013]. The convergence rates of the last iterate for online pairwise learning algorithms were studied based on convex analysis [Ying and Zhou, 2016; Guo *et al.*, 2016; Lin *et al.*, 2017], which, however, as we will show below, are suboptimal.

3 Problem Setup and Main Results

3.1 Regularized Pairwise Learning

Let ρ be a probability measure defined over the sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and output space, respectively. Let $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^n$ be a sequence of examples independently drawn from ρ. Let $\mathcal{W}$ be a Hilbert space with the associated inner product $\langle\cdot,\cdot\rangle$ and $\|\cdot\|, |||\cdot|||$ be two norms defined in $\mathcal{W}$. Let $\phi : \mathcal{X} \times \mathcal{X} \to \mathcal{W}$ be a feature map and $\tau : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The definition of $\tau(y, \tilde{y})$ depends on the specific application domain (See Examples 1, 2 below). From the feature map ϕ, one can define a Mercer kernel $K : (\mathcal{X} \times \mathcal{X}) \times (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}$ satisfying $K((x_1, \tilde{x}_1), (x_2, \tilde{x}_2)) = \langle\phi(x_1, \tilde{x}_1), \phi(x_2, \tilde{x}_2)\rangle$ for the reproducing Kernel Hilbert space $\mathcal{W}$ [Clémençon *et al.*, 2008; Rejchel, 2012]. For any two examples $z = (x, y), \tilde{z} = (\tilde{x}, \tilde{y})$ and a prediction rule $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we use the loss function of the form $V(h, z, \tilde{z}) = \ell(\tau(y, \tilde{y}), h(x, \tilde{x}))$ to measure the quality of h on z and $\tilde{z}$, where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is convex with respect to (w.r.t.) the second argument. We assume V is symmetric in the sense that $V(h, z, \tilde{z}) = V(h, \tilde{z}, z)$. The generalization error of any $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as $\mathcal{E}(h) := \mathbb{E}_{z,\tilde{z}}[V(h, z, \tilde{z})]$, where $\mathbb{E}_{z,\tilde{z}}$ denotes the conditional expectation w.r.t. z and $\tilde{z}$. We omit the subscript $z, \tilde{z}$ if the expectation is taken over all random variables. We consider prediction rules of the form $h_{\mathbf{w}}(x, \tilde{x}) := \langle\mathbf{w}, \phi(x, \tilde{x})\rangle, \mathbf{w} \in \mathcal{W}$, and we search the estimator $\mathbf{w}_{\mathbf{z},\lambda}$ by minimizing the empirical error plus a regularization term to avoid overfitting

$$\mathbf{w}_{\mathbf{z},\lambda} = \arg\min_{\mathbf{w}\in\mathcal{W}} \Big[F_{\mathbf{z},\lambda}(\mathbf{w}) := \lambda|||\mathbf{w}|||^2 + \frac{1}{n(n-1)} \sum_{i,j\in\mathbb{N}_n, i\neq j} \ell(\tau(y_i, y_j), \langle\mathbf{w}, \phi(x_i, x_j)\rangle)\Big], \quad (1)$$

where $\lambda > 0$ is a regularization parameter and we use the notation $\mathbb{N}_n = \{1, \ldots, n\}$. The regularized generalization error of the prediction rule $h_{\mathbf{w}}$ is defined as

$$F_\lambda(\mathbf{w}) = \mathbb{E}_{z,\tilde{z}}[\ell(\tau(y, \tilde{y}), \langle\mathbf{w}, \phi(x, \tilde{x})\rangle)] + \lambda|||\mathbf{w}|||^2. \quad (2)$$

We denote by $\mathbf{w}_\lambda$ the model minimizing $F_\lambda(\mathbf{w})$ over $\mathcal{W}$ and by h_ρ the model minimizing the generalization error over measurable functions defined on $\mathcal{X} \times \mathcal{X}$, respectively

$$\mathbf{w}_\lambda := \arg\min_{\mathbf{w}\in\mathcal{W}} F_\lambda(\mathbf{w}) \quad \text{and} \quad h_\rho := \arg\min_h \mathcal{E}(h). \quad (3)$$

The framework of pairwise learning covers many machine learning problems as specific examples. Here we clarify how the distance metric learning and ranking can be recovered with specific instantiations of ℓ, ϕ and τ. We denote $(t)_+ := \max(t, 0)$ and $x^\top$ the transpose of $x \in \mathbb{R}^d$.

Example 1 (Supervised metric learning)**.** Assume $\mathcal{Y} = \{\pm 1\}$. Supervised metric learning aims to find a Mahalanobis metric $d_{\mathbf{w}}(x_i, x_j) = \langle\mathbf{w}, (x_i - x_j)(x_i - x_j)^\top\rangle, \mathbf{w} \in \mathbb{S}^{d\times d}$ such that two examples with the same label are close to each other, while two examples with different labels are apart from each other. A common loss function used in metric learning takes the form $V_m(d_{\mathbf{w}}, z, \tilde{z}) = g(y\tilde{y}(1 - d_{\mathbf{w}}(x, \tilde{x})))$ [Jin *et al.*, 2009], where $g : \mathbb{R} \to \mathbb{R}_+$ is a convex function for which a typical choice is $g(t) = (1 - t)_+$. This scheme falls into the pairwise learning framework if we define $\tau(y, \tilde{y}) = y\tilde{y}, \phi(x, \tilde{x}) = (x - \tilde{x})(x - \tilde{x})^\top, \ell(a, b) = g(a(1 - b))$ and $h_{\mathbf{w}}(x, \tilde{x}) = \langle\mathbf{w}, \phi(x, \tilde{x})\rangle, \mathbf{w} \in \mathbb{S}^{d\times d}$. That is, we have $V_m(d_{\mathbf{w}}, z, \tilde{z}) = \ell(\tau(y, \tilde{y}), \langle\mathbf{w}, \phi(x, \tilde{x})\rangle)$.

Example 2 (Ranking)**.** In ranking problems, we use the output y_i to indicate the ordering between instances, i.e., the instance x_i is considered to be better than x_j if $y_i > y_j$. The task is to predict the ordering between the objects based on observations by constructing ranking rules $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and predict $y > \tilde{y}$ if $h(x, \tilde{x}) > 0$ [Rejchel, 2012; Clémençon *et al.*, 2008]. A common pairwise loss function used in ranking problems takes the form $V_r(h, z, \tilde{z}) = g(\text{sign}(y - \tilde{y})h(x, \tilde{x}))$, where $g : \mathbb{R} \to \mathbb{R}_+$ is a convex function which can be either the exponential cost function $g(t) = e^{-t}$, the "logit" function $g(t) = \log(1 + e^{-t})$ or the "hinge loss" $g(t) = (1 - t)_+$ [Clémençon *et al.*, 2008]. Here $\text{sign}(t)$ denotes the sign of $t \in \mathbb{R}$. The above formulation of ranking problems falls into our framework of pairwise learning by taking $\tau(y, \tilde{y}) = \text{sign}(y - \tilde{y}), \ell(a, b) = g(ab)$ and ranking rules of the form $h_{\mathbf{w}}(x, \tilde{x}) = \langle\mathbf{w}, \phi(x, \tilde{x})\rangle, \mathbf{w} \in \mathcal{W}$. That is, we have $V_r(h_{\mathbf{w}}, z, \tilde{z}) = \ell(\tau(y, \tilde{y}), \langle\mathbf{w}, \phi(x, \tilde{x})\rangle)$.

3.2 Main Results

We now present our main result, showing that the uniform deviation of the population sub-optimality $F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda)$ from the empirical sub-optimality $F_{\mathbf{z},\lambda}(\mathbf{w}) - F_{\mathbf{z},\lambda}(\mathbf{w}_\lambda)$ decays with the rate $O(1/n)$, provided that F_λ is strongly convex w.r.t. the norm $\|\cdot\|$, the dual norm of which is denoted by $\|\cdot\|_*$. Introduce the notation $X_* = \sup_{x,\tilde{x}\in\mathcal{X}} \|\phi(x,\tilde{x})\|_*$. For any $\mathbf{w} \in \mathcal{W}$, we refer to $F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda)$ as the excess regularized generalization error (ERGE) of $\mathbf{w}$. Let $e = \exp(1)$.

Definition 1 (Strong Convexity)**.** A function $f : \mathcal{W} \to \mathbb{R}$ is said to be β-strongly convex ($\beta > 0$) w.r.t. $\|\cdot\|$ if $\forall \mathbf{w}, \tilde{\mathbf{w}} \in \mathcal{W}$ and $\forall \alpha \in (0,1)$, we have $f(\alpha\mathbf{w} + (1-\alpha)\tilde{\mathbf{w}}) \leq \alpha f(\mathbf{w}) + (1-\alpha)f(\tilde{\mathbf{w}}) - \frac{\beta}{2}\alpha(1-\alpha)\|\mathbf{w} - \tilde{\mathbf{w}}\|^2$.

Theorem 1 (Main theorem)**.** *Let $L > 0$ and $\beta > 0$. Assume the loss function ℓ is L-Lipschitz continuous in the sense*

$$|\ell(\tau(y,\tilde{y}),a) - \ell(\tau(y,\tilde{y}),b)| \leq L|a-b|,\ \forall y,\tilde{y} \in \mathcal{Y}, a,b \in \mathbb{R}. \tag{4}$$

Assume that $F_\lambda(\mathbf{w})$ defined in Eq. (2) *is β-strongly convex w.r.t. the norm $\|\cdot\|$. Let $\mathbf{w}_\lambda$ be defined as Eq.*(3) *and any $\rho_0 > 0$. Then, for any $0 < \delta_0 < 1/e$, with probability at least $1 - 2\delta_0$ the following inequality holds for all $\mathbf{w} \in \mathcal{W}$*

$$F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda) \leq F_{\mathbf{z},\lambda}(\mathbf{w}) - F_{\mathbf{z},\lambda}(\mathbf{w}_\lambda) + \Big(1 + \sqrt{\log\delta_0^{-1} + \log\max(1, 2\rho_0^{-1}(F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda)))}\Big) \times 4LX_* \sqrt{\frac{2}{n\beta}\max(\rho_0, 2(F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda)))}. \tag{5}$$

In particular, for $\mathbf{w}_{\mathbf{z},\lambda}$ in Eq.(1)*, we have the following inequality with probability at least $1 - 2\delta_0$ for all $0 < a < 1$*

$$F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda) \leq \frac{128L^2X_*^2}{n\beta(1-a)}\log\frac{1}{a\delta_0}. \tag{6}$$

Note the norm $|||\cdot|||$ in the definition of regularization algorithm (1) is not necessarily equal to the norm $\|\cdot\|$ w.r.t. which F_λ is strongly convex. In learning theory, we often refer to the term $D(\lambda) := \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_\lambda|||^2$ as the approximation error. It is a standard assumption that the approximation error admits a polynomial decay rate [Smale and Zhou, 2003; Ying and Zhou, 2016; Zhao *et al.*, 2017]. In particular, if $h_\rho \in \mathcal{W}$ then $D(\lambda) \leq \lambda|||h_\rho|||^2$ [Guo *et al.*, 2016]. Throughout the paper we use the abbreviation $\mathcal{E}(\mathbf{w}) := \mathcal{E}(h_\mathbf{w})$.

Assumption 1. We assume the approximation error $D(\lambda)$ enjoys a polynomial decay rate with exponent $0 < \alpha \leq 1$ in the sense $D(\lambda) \leq c_\alpha\lambda^\alpha, \forall\lambda > 0$, where $c_\alpha > 0$ is a constant.

Theorem 2. *Suppose that the assumptions in Theorem 1 and Assumption 1 hold. Then, for any $0 < \delta_0 < 1/e$, with probability $1 - 2\delta_0$ there holds*

$$\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_{\mathbf{z},\lambda}|||^2 \leq c_\alpha\lambda^\alpha + \frac{256L^2X_*^2}{n\beta}\log\frac{2}{\delta_0}.$$

Proof. Plugging $a = \frac{1}{2}$ in Eq. (6) and recalling $F_\lambda(\mathbf{w}) = \mathcal{E}(\mathbf{w}) + \lambda|||\mathbf{w}|||^2$, with probability at least $1 - 2\delta_0$ we can upper bound the term $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_{\mathbf{z},\lambda}|||^2$ by

$$\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_\lambda|||^2 + \frac{256L^2X_*^2}{n\beta}\log\frac{2}{\delta_0}.$$

The stated bound can be derived by plugging Assumption 1 into the above inequality. □

Remark 1. The regularization parameter λ should be chosen according to β and n to achieve optimal generalization bounds. Convergence rates $O(\log(n)/(n\beta^2))$ were established for the n-th iteration of online regularized pairwise learning algorithms [Guo *et al.*, 2016], which are suboptimal to (6) if $\beta \ll 1$. For example, under conditions of Theorem 2 with $\beta \asymp \lambda$, Theorem 2 with $\lambda \asymp n^{-\frac{1}{\alpha+1}}$ implies the learning rate $O(n^{-\frac{\alpha}{\alpha+1}})$ with high probability, while the discussions in [Guo *et al.*, 2016] and [Lin *et al.*, 2017] can only imply the rate $O(n^{-\frac{\alpha}{2(\alpha+1)}}\log(n))$ and the rate $O(n^{-\frac{\alpha}{1+2\alpha}}\log(n))$ in expectation, respectively. Here, $A \asymp B$ means there are universal constants $C_1, C_2 > 0$ with $C_1A \leq B \leq C_2A$. In Section 4, we will apply Theorem 2 to improve the existing bounds for metric learning and ranking.

Remark 2. Another term of interest in the literature of regularized pairwise learning is $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}_\mathbf{z}(\mathbf{w}_{z,\lambda})$. We now relate this to $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_{\mathbf{z},\lambda}|||^2$. On the one hand, by Theorem 8.3 in [Cucker and Zhou, 2007] and $|||\mathbf{w}_\lambda||| \leq \sqrt{D(\lambda)/\lambda}$, we have $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_{\mathbf{z},\lambda}|||^2 = O(\lambda^{\frac{\alpha-1}{2}}n^{-\frac{1}{2}}) + \mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}_\mathbf{z}(\mathbf{w}_{\mathbf{z},\lambda}) + D(\lambda)$ with high probability, where $\mathcal{E}_\mathbf{z}(\mathbf{w}) = F_{\mathbf{z},\lambda}(\mathbf{w}) - \lambda|||\mathbf{w}|||^2$ is empirical error (without regularizer). On the other hand, we can take $\mathbf{w} = \mathbf{w}_{\mathbf{z},\lambda}$ in (5) to get $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}_\mathbf{z}(\mathbf{w}_{\mathbf{z},\lambda}) \leq \mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}_\mathbf{z}(\mathbf{w}_\lambda) + E$ with high probability, where E is last term in (5) with $\mathbf{w} = \mathbf{w}_{\mathbf{z},\lambda}$. Taking ρ_0 in (5) as the right-hand side of (6) and using (6) we get $E = O(1/(n\beta))$. This, together with $\mathcal{E}(\mathbf{w}_\lambda) - \mathcal{E}_\mathbf{z}(\mathbf{w}_\lambda) = O(\lambda^{\frac{\alpha-1}{2}}n^{-\frac{1}{2}})$, allows us to derive $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}_\mathbf{z}(\mathbf{w}_{\mathbf{z},\lambda}) = O\big(\lambda^{\frac{\alpha-1}{2}}n^{-\frac{1}{2}} + 1/(n\beta)\big)$, which improves the bound $O(1/\sqrt{n\lambda})$ based on the U-process approach [Cao *et al.*, 2016] and the bound $O(1/(\lambda\sqrt{n}))$ based on the stability approach since we often have $1/n \ll \beta \asymp \lambda \ll 1$ (for example, Theorem 2 suggests $\lambda \asymp n^{-\frac{1}{\alpha+1}}$).

4 Applications

We now apply Theorem 2 to distance metric learning and ranking. We always assume $g : \mathbb{R} \to \mathbb{R}_+$ is convex.

4.1 Regularized Distance Metric Learning

As clarified in Example 1, an established regularization framework to learn the Mahalanobis distance metric can be reformulated as a regularized pairwise learning problem (1) with some specific instantiations of $\mathcal{W}, \tau, \phi$ and ℓ. Some interesting regularizers in metric learning include the ℓ_1-norm regularizer favoring the element-wise sparsity [Cao *et al.*, 2016], the mixed $(2,1)$-norm regularizer encouraging the column-wise sparsity [Ying *et al.*, 2009] and the trace-norm regularizer encouraging low-rank [Cao *et al.*, 2016], etc. For any $\mathbf{w} \in \mathbb{R}^{d\times d}$, we define the Schatten-p norm $\|\mathbf{w}\|_{S(p)}$ as the ℓ_p-norm of $\sigma(\mathbf{w}), p \geq 1$, where $\sigma(\mathbf{w})$ is the vector of all singular values of $\mathbf{w}$ and the ℓ_p-norm of $\mathbf{a} = (a_1, \ldots, a_d) \in \mathbb{R}^d$ is defined as $\|\mathbf{a}\|_p = \big[\sum_{j=1}^d |a_j|^p\big]^{1/p}$. For any $\mathbf{w} = (\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^m) \in \mathbb{R}^{n\times m}$, the mixed (p,q)-norm of $\mathbf{w}$

is $\|\mathbf{w}\|_{(p,q)} := \big\|\big(\|\mathbf{w}^1\|_p, \ldots, \|\mathbf{w}^m\|_p\big)\big\|_q, \forall p, q \geq 1$. It is known that the dual norm of $\|\cdot\|_{p,q}$ is $\|\cdot\|_{(p^*,q^*)}$, and the dual norm of $\|\cdot\|_{S(p)}$ is $\|\cdot\|_{S(p^*)}$, where p^* is the conjugate exponent of p satisfying $p^{-1} + (p^*)^{-1} = 1$. For brevity, we introduce some notations $X_{(p,q)} = \sup_{x,\bar{x}\in\mathcal{X}} \|(x-\bar{x})(x-\bar{x})^\top\|_{(p,q)}$, $X_{S(p)} = \sup_{x,\bar{x}\in\mathcal{X}} \|(x-\bar{x})(x-\bar{x})^\top\|_{S(p)}$, $X_p = \sup_{x,\bar{x}\in\mathcal{X}} \|x-\bar{x}\|_p$. Denote $\tilde{d} = \frac{\log d}{\log d - 1}$.

We now apply Theorem 2 to study regularized metric learning with the regularizer involving various norms, including the mixed (p,q)-norm, the mixed $(p,\tilde{d})$-norm, the $S(\tilde{d})$-norm, and the $S(p)$-norm. Since $\|\cdot\|^2_{(p,1)}, \|\cdot\|^2_{S(1)}$ are not strongly convex, we use the strongly convex regularizers $\|\cdot\|^2_{(p,\tilde{d})}$ and $\|\cdot\|^2_{S(\tilde{d})}$ [Kakade *et al.*, 2012] to mimic the effects of the mixed $(p,1)$-norm based regularizer and the Schatten-1 norm based regularizer, respectively. We hide the linear dependency on $\log \delta_0^{-1}$ in the big O notation.

Corollary 3. *Consider the regularized metric learning* (1) *with* $\mathcal{W} = \mathbb{S}^{d\times d}, \tau(y,\tilde{y}) = y\tilde{y}, \phi(x,\tilde{x}) = (x-\tilde{x})(x-\tilde{x})^\top$ *and* $\ell(a,b) = g(a(1-b))$. *Let* $p, q \in (1,2]$ *and* $0 < \delta_0 < 1/e$. *Assume* (4) *and Assumption 1 hold. If we choose* $\lambda \asymp X_*^{\frac{2}{\alpha+1}} n^{-\frac{1}{\alpha+1}}$, *then with probability* $1-2\delta_0$ *we have* $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_{\mathbf{z},\lambda}|||^2 =$

$$\begin{cases} O((X_{p^*}X_{q^*})^{\frac{2\alpha}{\alpha+1}} n^{-\frac{\alpha}{\alpha+1}}), & \text{if } |||\cdot||| = \|\cdot\|_{(p,q)}, \text{ (7a)} \\ O((X^2_{p^*}X^2_\infty \log d)^{\frac{\alpha}{\alpha+1}} n^{-\frac{\alpha}{\alpha+1}}), & \text{if } |||\cdot||| = \|\cdot\|_{(p,\tilde{d})}, \text{ (7b)} \\ O(X_{S(p^*)}^{\frac{2\alpha}{\alpha+1}} n^{-\frac{\alpha}{\alpha+1}}), & \text{if } |||\cdot||| = \|\cdot\|_{S(p)}, \text{ (7c)} \\ O((X^2_{S(\infty)} \log d)^{\frac{\alpha}{\alpha+1}} n^{-\frac{\alpha}{\alpha+1}}), & \text{if } |||\cdot||| = \|\cdot\|_{S(\tilde{d})}. \text{ (7d)} \end{cases}$$

Proof. It was shown in [Kakade *et al.*, 2012] that $r(M) := \lambda\|M\|^2_{(p,q)}$ is $2\lambda\frac{(p-1)(q-1)}{p+q-2}$-strongly convex w.r.t. the norm $\|\cdot\|_{(p,q)}$. Therefore, Theorem 2 shows that, with probability $1-2\delta_0$, the term $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda\|\mathbf{w}_{\mathbf{z},\lambda}\|^2_{(p,q)}$ can be upper bounded by

$$\frac{128L^2X^2_{(p^*,q^*)}(p+q-2)}{\lambda n(p-1)(q-1)} \log\frac{2}{\delta_0} + c_\alpha\lambda^\alpha.$$

Eq. (7a) then follows by taking $\lambda \asymp (X_{p^*}X_{q^*})^{\frac{2}{\alpha+1}} n^{-\frac{1}{\alpha+1}}$ and noticing $X_{(p^*,q^*)} = X_{p^*}X_{q^*}$.

Eqs. (7b), (7c) and (7d) can be derived in a similar way. We omit the deduction for brevity. □

Remark 3 (Comparison with the state of the art)**.** We now compare our learning rates to the state-of-the-art bounds for regularized metric learning. We complement the existing estimation error bounds with approximation error bounds under Assumption 1. Jin *et al.* [2009] studied regularized distance metric learning (1) with $\tau(y,\tilde{y}) = y\tilde{y}, \phi(x,\tilde{x}) = (x-\tilde{x})(x-\tilde{x})^\top$ and $|||\cdot|||$ being the Frobenius norm, and derived the error bound $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_{\mathbf{z},\lambda}|||^2 = O\Big(\big(\frac{X_2^4}{\lambda} + \sqrt{\frac{d}{\lambda}}\big)\sqrt{\frac{\log\delta_0^{-1}}{n}} + \lambda^\alpha\Big)$. Taking $\lambda \asymp X_2^{\frac{4}{\alpha+1}} n^{-\frac{1}{2(\alpha+1)}}$, this bound becomes $O\Big(X_2^{\frac{4\alpha}{\alpha+1}} n^{-\frac{\alpha}{2(\alpha+1)}} + \sqrt{d} n^{-\frac{2\alpha+1}{4(\alpha+1)}}\Big)$, which is worse than the bound $O\big(X_2^4 n^{-1}\big)^{\frac{\alpha}{\alpha+1}}$ in (7a) with $p=q=2$. For regularized metric learning (1) with a general regularizer, with probability $1-\delta_0$ the term $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_{\mathbf{z},\lambda}|||^2$ was upper bounded in [Cao *et al.*, 2016] by

$$c_\alpha\lambda^\alpha + O\Bigg(\frac{\mathbb{E}_{\mathbf{z},\sigma} \sup_{|||\mathbf{w}|||\leq 1} \sum_{i=1}^n \sigma_i \langle \mathbf{w}, x_i x^\top_{\lfloor \frac{n}{2}\rfloor + i}\rangle}{\sqrt{\lambda}n}\Bigg), \quad (8)$$

where $\{\sigma_i\}_{i\in\mathbb{N}_n}$ is a sequence of independent Rademacher variables taking the value $+1$ or -1 with equal probability. The term $\mathbb{E}_{\mathbf{z},\sigma}\sup_{|||\mathbf{w}|||\leq 1}\sum_{i=1}^n \sigma_i\langle \mathbf{w}, x_i x^\top_{\lfloor \frac{n}{2}\rfloor+i}\rangle$ is closely related to the *Rademacher complexity for metric learning* in [Cao *et al.*, 2016] and can be estimated by the seminal complexity bound of linear prediction with specific instantiations of strongly convex functions [Kakade *et al.*, 2012]. In Table 1, we list the best generalization error bounds derived by choosing appropriate regularization parameters λ to minimize (8). It is clear from Table 1 that our generalization analysis yields the bounds $O(n^{-\frac{\alpha}{\alpha+1}})$, which significantly improve the bound $O(n^{-\frac{\alpha}{2\alpha+1}})$ in [Cao *et al.*, 2016].

4.2 Regularized Ranking

As shown in Example 2, a regularized ranking problem is a specific regularized pairwise learning problem with some specific instantiations of τ and ℓ. In this subsection, we assume both $\|\cdot\|$ and $|||\cdot|||$ are the norm induced by the inner product $\langle\cdot,\cdot\rangle$ in $\mathcal{W}$, i.e., $\|\mathbf{w}\| = \sqrt{\langle\mathbf{w},\mathbf{w}\rangle}$ for any $\mathbf{w}\in\mathcal{W}$. The following corollary follows directly from Theorem 2 by noting the (2λ)-strong convexity of $\lambda\|\cdot\|^2$ w.r.t. $\|\cdot\|$. We omit the proof due to the space constraint.

Corollary 4. *Consider the regularized ranking* (1) *with* $\ell(a,b) = g(ab)$ *and either* $\tau(y,\tilde{y}) = sign(y-\tilde{y})$ *or* $\tau(y,\tilde{y}) = y-\tilde{y}$. *Assume* (4) *and Assumption 1 hold. If we choose* $\lambda \asymp X_*^{\frac{2}{\alpha+1}} n^{-\frac{1}{\alpha+1}}$, *then with probability* $1-2\delta_0$ *we have* $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_{\mathbf{z},\lambda}|||^2 = O\big((X_*^2 n^{-1})^{\frac{\alpha}{\alpha+1}}\big)$.

Remark 4 (Comparison with the state of the art)**.** Fast learning rates were established for unregularized ranking in a compact hypothesis space satisfying a capacity assumption with the loss function satisfying an assumption on the modulus of convexity [Rejchel, 2012]. These error bounds are stated only for the estimation error (ignoring the approximation error), and can be as slow as $O(n^{-\frac{1}{2}})$ if the capacity assumption is removed. Agarwal and Niyogi [2009] studied the generalization performance of regularized ranking algorithms based on a stability approach, and derived the bound $\mathcal{E}(\mathbf{w}_{\mathbf{z},\lambda}) - \mathcal{E}(h_\rho) + \lambda|||\mathbf{w}_{\mathbf{z},\lambda}|||^2 = O\big(\frac{1}{\lambda\sqrt{n}} + \lambda^\alpha\big)$. If we take $\lambda \asymp n^{-\frac{1}{2(\alpha+1)}}$, then this bound becomes $O\big(n^{-\frac{\alpha}{2(\alpha+1)}}\big)$, which is improved to $O(n^{-\frac{\alpha}{\alpha+1}})$ in Corollary 4 without capacity assumptions on the hypothesis spaces. Zhao *et al.* [2017] derived generalization bounds $O(n^{\epsilon-\frac{\alpha}{1+\alpha}})$ for a specific regularized ranking problem (1) with $\tau(y,\tilde{y}) = y-\tilde{y}$ and $\ell(a,b) = (a-b)^2$. Here ϵ is a positive constant. Therefore, even for this specific regularized ranking problem, our general error bound for regularized pairwise learning (1) is able to imply a tighter bound than existing bounds derived

	$\|\cdot\|_{(p,q)}, p,q \in (1,2]$	$\|\cdot\|_{(p,\tilde{d})}, p \in (1,2]$	$\|\cdot\|_{S(p)}, p \in (1,2]$	$\|\cdot\|_{S(\tilde{d})}$
Ours	$(X_{p^*}^2 X_{q^*}^2 n^{-1})^{\frac{\alpha}{\alpha+1}}$	$(X_{p^*}^2 X_{\infty}^2 (\log d) n^{-1})^{\frac{\alpha}{\alpha+1}}$	$(X_{S(p^*)}^2 n^{-1})^{\frac{\alpha}{\alpha+1}}$	$(X_{S(\infty)}^2 (\log d) n^{-1})^{\frac{\alpha}{\alpha+1}}$
[Cao *et al.*, 2016]	$(X_{p^*}^2 X_{q^*}^2 n^{-1})^{\frac{\alpha}{2\alpha+1}}$	$(X_{p^*}^2 X_{\infty}^2 (\log d) n^{-1})^{\frac{\alpha}{2\alpha+1}}$	$(X_{S(p^*)}^2 n^{-1})^{\frac{\alpha}{2\alpha+1}}$	$(X_{S(\infty)}^2 (\log d) n^{-1})^{\frac{\alpha}{2\alpha+1}}$

Table 1: Comparison of generalization bounds for regularized metric learning. The first row shows different instantiations of the norm $|||\cdot|||$ in (1), for which the related generalization bounds established in this paper and [Cao *et al.*, 2016] are presented in the second and third row.

exclusively for a specific regularized ranking algorithm using the least squares loss. It should be mentioned that the bound $O(n^{\epsilon-\frac{\alpha}{1+\alpha}})$ in [Zhao *et al.*, 2017] can be improved if a further capacity assumption on the hypothesis space is imposed.

5 Proof of the Main Theorem

In this section, we present the proof of the main theorem (Theorem 1). The key idea in our deduction is to partition the hypothesis space into a sequence of subsets according to the value of ERGE, using the idea of peeling [Bartlett *et al.*, 2005]. The strong convexity of the regularized objective then implies that the infinity-norm of functions in each sub-class can be bounded by the maximal ERGE associated to that sub-class, which allows us to conduct the localization analysis to control the uniform deviation between the ERGE and its empirical counterpart in each sub-class by the local ERGE. For any $\mathbf{w} \in \mathcal{W}$, define the excess loss function by

$$h_{\mathbf{w}}^{\lambda}(z,\tilde{z}) := \ell(\tau(y,\tilde{y}), \langle \mathbf{w}, \phi(x,\tilde{x})\rangle) - \ell(\tau(y,\tilde{y}), \langle \mathbf{w}_\lambda, \phi(x,\tilde{x})\rangle).$$

Let $\rho_0 > 0$ and $0 < \delta_0 < 1/e$ be any two fixed number. We construct two geometric sequences $\rho_k = 2^k \rho_0, \delta_k = 2^{-k}\delta_0, k = 1, 2, \ldots$. For brevity, we set $\rho_{-1} = 0$. We group those $\mathbf{w}$ whose ERGEs belong to $(\rho_{k-1}, \rho_k]$ into the class $\mathcal{W}_k$

$$\mathcal{W}_k = \{\mathbf{w} \in \mathcal{W} : \rho_{k-1} < F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda) \le \rho_k\},\ k \in \mathbb{N} \cup \{0\}.$$

Define $\mathcal{H}_k = \{h_{\mathbf{w}}^{\lambda} : \mathbf{w} \in \mathcal{W}_k\}, k \in \mathbb{N} \cup \{0\}$.

We denote by $\lfloor x \rfloor$ the largest natural number not larger than x. We begin our deduction with the following lemma controlling the infinity-norm of functions in $\mathcal{H}_k$.

Lemma 5. *If* (4) *holds and F_λ is β-strongly convex, then* $\sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} \|h_{\mathbf{w}}^{\lambda}\|_\infty \le 2LX_* \sqrt{\frac{\rho_k}{\beta}}$.

Proof. For any $h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k$, $\|h_{\mathbf{w}}^{\lambda}\|_\infty$ can be upper bounded by

$$\begin{aligned} &\sup_{z,\tilde{z}} |\ell(\tau(y,\tilde{y}), \langle \mathbf{w}, \phi(x,\tilde{x})\rangle) - \ell(\tau(y,\tilde{y}), \langle \mathbf{w}_\lambda, \phi(x,\tilde{x})\rangle)| \\ &\le \sup_{z,\tilde{z}} L |\langle \mathbf{w} - \mathbf{w}_\lambda, \phi(x,\tilde{x})\rangle| \le L X_* \|\mathbf{w} - \mathbf{w}_\lambda\|, \end{aligned} \tag{9}$$

where the first inequality follows from the Lipschitz property of the loss function ℓ and the second inequality is due to the definition of dual norm together with the definition of X_*. According to the definition of $\mathbf{w}_\lambda$ and the strong convexity of F_λ, the following inequality holds for all $\mathbf{w} \in \mathcal{W}$

$$F_\lambda(\mathbf{w}_\lambda) \le F_\lambda\Big(\frac{\mathbf{w}+\mathbf{w}_\lambda}{2}\Big) \le \frac{F_\lambda(\mathbf{w})}{2} + \frac{F_\lambda(\mathbf{w}_\lambda)}{2} - \frac{\beta\|\mathbf{w}-\mathbf{w}_\lambda\|^2}{8},$$

which, together with the assumption $h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k$, implies

$$\|\mathbf{w} - \mathbf{w}_\lambda\| \le \sqrt{\frac{4}{\beta}(F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda))} \le \sqrt{\frac{4\rho_k}{\beta}},\ \forall \mathbf{w} \in \mathcal{W}_k.$$

Putting this inequality into (9) and using the definition of $h_{\mathbf{w}}^{\lambda}$ show $\|h_{\mathbf{w}}^{\lambda}\|_\infty \le \sqrt{\frac{4\rho_k}{\beta}} L X_*$. □

We prove Theorem 1 in four steps. For any f defined over $\mathcal{Z} \times \mathcal{Z}$, we use $\hat{\mathbb{E}}_{\mathbf{z}} f = \frac{1}{n(n-1)} \sum_{i,j \in \mathbb{N}_n, i \ne j} f(z_i, z_j)$ to denote the empirical average of f. In the first and second step, we conduct localization analysis in each sub-class $\mathcal{H}_k$ and get a bound with probability $1-\delta_k$ for $\sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} [\mathbb{E}[h_{\mathbf{w}}^{\lambda}] - \hat{\mathbb{E}}_{\mathbf{z}}[h_{\mathbf{w}}^{\lambda}]]$ in terms of ρ_k. In the third step, we show both δ_k^{-1} and ρ_k can be bounded by ERGE, which together with union bounds, implies a probability inequality on ERGE in terms of a square root of ERGE. In the last step, we solve this probabilistic inequality to get the stated bound. The intuitive observation is that both the reciprocal of confidence and the constant in the bounded difference property for the application of McDiarmid's inequality in each sub-class can be controlled by ERGE associated to that sub-class, which allows us to get bounds on ERGE in terms of a square root of ERGE.

Proof of Theorem 1. Our proof consists of four steps.

Step 1. We first apply the McDiarmid's inequality (Theorem D.3 in [Mohri *et al.*, 2012]) to control the deviation of $\sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} [\mathbb{E}[h_{\mathbf{w}}^{\lambda}] - \hat{\mathbb{E}}_{\mathbf{z}}[h_{\mathbf{w}}^{\lambda}]]$ from its expectation. To this aim, we need to show that functions in $\mathcal{H}_k$ satisfy the bounded difference property. Indeed, for any $\mathbf{z} = \{z_1, \ldots, z_{t-1}, z_t, z_{t+1}, \ldots, z_n\}$ and $\bar{\mathbf{z}} = \{z_1, \ldots, z_{t-1}, \bar{z}_t, z_{t+1}, \ldots, z_n\}$, we have

$$\begin{aligned} &\Big| \sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} (\mathbb{E}[h_{\mathbf{w}}^{\lambda}] - \hat{\mathbb{E}}_{\mathbf{z}}[h_{\mathbf{w}}^{\lambda}]) - \sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} (\mathbb{E}[h_{\mathbf{w}}^{\lambda}] - \hat{\mathbb{E}}_{\bar{\mathbf{z}}}[h_{\mathbf{w}}^{\lambda}]) \Big| \\ &\le \sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} |\hat{\mathbb{E}}_{\mathbf{z}}[h_{\mathbf{w}}^{\lambda}] - \hat{\mathbb{E}}_{\bar{\mathbf{z}}}[h_{\mathbf{w}}^{\lambda}]| \\ &\le \frac{2}{n(n-1)} \sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} \sum_{j \in \mathbb{N}_n, j \ne t} (|h_{\mathbf{w}}^{\lambda}(z_t, z_j)| + |h_{\mathbf{w}}^{\lambda}(\bar{z}_t, z_j)|) \\ &\le \frac{4}{n} \sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} \|h_{\mathbf{w}}^{\lambda}\|_\infty \le \frac{8LX_*}{n} \sqrt{\frac{\rho_k}{\beta}}, \end{aligned}$$

where the last inequality follows from Lemma 5. Applying McDiarmid's inequality (Theorem D.3 in [Mohri *et al.*, 2012]) with increments bounded by $\frac{8LX_*}{n}\sqrt{\frac{\rho_k}{\beta}}$ implies that with probability at least $1-\delta_k$ the term $\sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} [\mathbb{E}[h_{\mathbf{w}}^{\lambda}] - \hat{\mathbb{E}}_{\mathbf{z}}[h_{\mathbf{w}}^{\lambda}]]$ can be upper bounded by

$$\mathbb{E} \sup_{h_{\mathbf{w}}^{\lambda} \in \mathcal{H}_k} [\mathbb{E}[h_{\mathbf{w}}^{\lambda}] - \hat{\mathbb{E}}_{\mathbf{z}}[h_{\mathbf{w}}^{\lambda}]] + 4LX_* \sqrt{\frac{2\rho_k \log(1/\delta_k)}{n\beta}}. \tag{10}$$

Step 2. We now use techniques in U-process to bound the expectation in (10). Specifically, Lemma A.1 in

[Clémençon *et al.*, 2008] with $q_{\mathbf{w}}(z_i, z_j) = \mathbb{E}[h^\lambda_{\mathbf{w}}] - h^\lambda_{\mathbf{w}}(z_i, z_j)$ and the index set $\mathcal{W}$ allows us to derive $\mathbb{E}_{\mathbf{z}} \sup_{h^\lambda_{\mathbf{w}} \in \mathcal{H}_k} [\mathbb{E}[h^\lambda_{\mathbf{w}}] - \hat{\mathbb{E}}_{\mathbf{z}}[h^\lambda_{\mathbf{w}}]] \leq \mathbb{E}_{\mathbf{z}} \sup_{h^\lambda_{\mathbf{w}} \in \mathcal{H}_k} \Big[\mathbb{E}[h^\lambda_{\mathbf{w}}] - \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} h^\lambda_{\mathbf{w}}(z_i, z_{\lfloor \frac{n}{2} \rfloor + i})\Big]$. Let $\bar{\mathbf{z}} = \{\bar{z}_1, \ldots, \bar{z}_n\}$ be i.i.d. samples independent of $\mathbf{z}$ and let $\{\sigma_i\}_{i \in \mathbb{N}_n}$ be a sequence of independent Rademacher variables. According to Jensen's inequality and a standard symmetrization technique, the term $\mathbb{E}_{\mathbf{z}} \sup_{h^\lambda_{\mathbf{w}} \in \mathcal{H}_k} [\mathbb{E}[h^\lambda_{\mathbf{w}}] - \hat{\mathbb{E}}_{\mathbf{z}}[h^\lambda_{\mathbf{w}}]]$ can be upper bounded by

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z},\bar{\mathbf{z}}} \sup_{h^\lambda_{\mathbf{w}} \in \mathcal{H}_k} \frac{1}{\lfloor \frac{n}{2} \rfloor} \Big[\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} h^\lambda_{\mathbf{w}}(\bar{z}_i, \bar{z}_{\lfloor \frac{n}{2} \rfloor + i}) - \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} h^\lambda_{\mathbf{w}}(z_i, z_{\lfloor \frac{n}{2} \rfloor + i}) \Big] \\
&= \mathbb{E}_{\mathbf{z},\bar{\mathbf{z}},\sigma} \sup_{h^\lambda_{\mathbf{w}} \in \mathcal{H}_k} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i [h^\lambda_{\mathbf{w}}(\bar{z}_i, \bar{z}_{\lfloor \frac{n}{2} \rfloor + i}) - h^\lambda_{\mathbf{w}}(z_i, z_{\lfloor \frac{n}{2} \rfloor + i})] \\
&\leq \frac{2}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z},\sigma} \sup_{\mathbf{w} \in \mathcal{W}_k} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \ell(\tau(y_i, y_{\lfloor \frac{n}{2} \rfloor + i}), \langle \mathbf{w}, \phi(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}) \rangle) \\
&\leq \frac{2}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_{\mathbf{z},\sigma} \sup_{\mathbf{w} \in \mathcal{W}_k} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i L \langle \mathbf{w}, \phi(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}) \rangle, \qquad (11)
\end{aligned}
$$

where the second inequality uses the fact that $\mathbf{w}_\lambda$ is a fixed element and the last inequality follows from a contraction property of Rademacher averages (Lemma 4.2 in [Mohri *et al.*, 2012] with the Lipschitz composition operator $t \mapsto \ell(\tau(y_i, y_{\lfloor \frac{n}{2} \rfloor + i}), t), i = 1, \ldots, \lfloor \frac{n}{2} \rfloor$).

The function $\bar{F}_\lambda(\mathbf{w}) := F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda)$ is β-strongly convex. Note that any $\mathbf{w} \in \mathcal{W}_k$ satisfies $\bar{F}_\lambda(\mathbf{w}) \leq \rho_k$. Moreover, the definition of $\mathbf{w}_\lambda$ implies that $\inf_{\mathbf{w} \in \mathcal{W}} \bar{F}_\lambda(w) = 0$. Therefore, the conditions of Theorem 7 in [Kakade *et al.*, 2012] are satisfied and we can apply it here to show

$$
\mathbb{E}_{\mathbf{z},\sigma} \sup_{\mathbf{w} \in \mathcal{W}_k} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \langle \mathbf{w}, \phi(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}) \rangle \leq X_* \sqrt{\frac{2 \lfloor \frac{n}{2} \rfloor \rho_k}{\beta}}.
$$

Plugging it back into Eq. (11) gives

$$
\mathbb{E}_{\mathbf{z}} \sup_{h^\lambda_{\mathbf{w}} \in \mathcal{H}_k} [\mathbb{E}[h^\lambda_{\mathbf{w}}] - \hat{\mathbb{E}}_{\mathbf{z}}[h^\lambda_{\mathbf{w}}]] \leq 2LX_* \sqrt{\frac{2\rho_k}{\lfloor \frac{n}{2} \rfloor \beta}}. \qquad (12)
$$

Step 3. We now present probabilistic bounds for $\sup_{\mathbf{w} \in \mathcal{W}} [\mathbb{E}[h^\lambda_{\mathbf{w}}] - \hat{\mathbb{E}}_{\mathbf{z}}[h^\lambda_{\mathbf{w}}]]$. For any $h^\lambda_{\mathbf{w}} \in \mathcal{H}_k, k \geq 1$, we have $\frac{\rho_k}{2} = \rho_{k-1} \leq F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda)$. Therefore,

$$
\rho_k \leq \max\{\rho_0, 2(F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda))\}, \ \forall \mathbf{w} \in \mathcal{W}_k, k \in \mathbb{N} \cup \{0\}.
$$

Moreover, it can be directly checked that

$$
\frac{1}{\delta_k} = \frac{\rho_0 2^k}{\delta_0 \rho_0} \leq \frac{\max\{\rho_0, 2(F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda))\}}{\delta_0 \rho_0}. \qquad (13)
$$

According to the definition of F_λ and $F_{\mathbf{z},\lambda}$ given in (1) and (2) together with the definition of $h^\lambda_{\mathbf{w}}$, we derive

$$
[F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda)] - [F_{\mathbf{z},\lambda}(\mathbf{w}) - F_{\mathbf{z},\lambda}(\mathbf{w}_\lambda)] = \mathbb{E}[h^\lambda_{\mathbf{w}}] - \hat{\mathbb{E}}_{\mathbf{z}}[h^\lambda_{\mathbf{w}}].
$$

Combining the above identity, (10), (12) and (13) together, with probability $1 - \delta_k$ the following inequality holds uniformly for all $h^\lambda_{\mathbf{w}} \in \mathcal{H}_k$

$$
\begin{aligned}
&F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda) \leq F_{\mathbf{z},\lambda}(\mathbf{w}) - F_{\mathbf{z},\lambda}(\mathbf{w}_\lambda) + \\
&\Big(1 + \sqrt{\log \delta_0^{-1} + \log \max(1, 2\rho_0^{-1}(F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda)))}\Big) \\
&\qquad \times 4LX_* \sqrt{\frac{2}{n\beta} \max(\rho_0, 2(F_\lambda(\mathbf{w}) - F_\lambda(\mathbf{w}_\lambda)))}.
\end{aligned}
$$

The stated inequality (5) follows directly if we apply union bounds over $\mathcal{H}_0, \mathcal{H}_1, \ldots$ with confidence $\delta_0, \delta_1, \ldots$ (notice that $\sum_{k=0}^\infty \delta_k = 2\delta_0$).

Step 4. Finally, we present probabilistic bounds for the estimator $\mathbf{w}_{\mathbf{z},\lambda}$. According to the definition of $\mathbf{w}_{\mathbf{z},\lambda}$, we have $F_{\mathbf{z},\lambda}(\mathbf{w}_{\mathbf{z},\lambda}) \leq F_{\mathbf{z},\lambda}(\mathbf{w}_\lambda)$ and therefore (5) implies

$$
\begin{aligned}
&\frac{F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda)}{4LX_*} \leq \sqrt{\frac{2\max(\rho_0, 2(F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda)))}{n\beta}} \\
&\times \Big(1 + \sqrt{\log \delta_0^{-1} + \log \max(1, 2\rho_0^{-1}(F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda)))}\Big).
\end{aligned}
$$

Take the assignment $\rho_0 = \frac{256L^2X_*^2}{n\beta}$. If $F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda) \geq \frac{\rho_0}{2}$, the term $F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda)$ can be further upper bounded by

$$
\frac{128L^2X_*^2}{n\beta}\Big(1 + \log \delta_0^{-1} + \log \frac{2(F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda))}{\rho_0}\Big).
$$

For the assignment $\rho_0 = \frac{256L^2X_*^2}{n\beta}$ and any $0 < a < 1$, the term $\frac{n\beta(F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda))}{128L^2X_*^2}$ can be upper bounded by (note $(a+b)^2 \leq 2(a^2 + b^2), \forall a, b \in \mathbb{R}$)

$$
\begin{aligned}
&1 + \log \delta_0^{-1} + \log \frac{n\beta(F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda))}{128L^2X_*^2} \\
&\leq \log \delta_0^{-1} + \frac{an\beta(F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda))}{128L^2X_*^2} + \log \frac{1}{a}, \qquad (14)
\end{aligned}
$$

where the last inequality follows from $\log a \leq ab + \log \frac{1}{b} - 1, \forall a, b > 0$. Solving the linear inequality (14) directly yields the stated bound (6) with probability at least $1 - 2\delta_0$. If $F_\lambda(\mathbf{w}_{\mathbf{z},\lambda}) - F_\lambda(\mathbf{w}_\lambda) \leq \frac{\rho_0}{2}$, it is clear that (6) holds. □

6 Conclusions

We develop a unified generalization bound for pairwise learning, and apply it to improve existing results. It would be interesting to study pairwise learning by exploiting the smoothness of loss functions [Zhang *et al.*, 2017] and consider distributed pairwise learning algorithms [Lin and Zhou, 2018].

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2017YFB1003102), the National Natural Science Foundation of China (Grant Nos. 61672478, 61502342), and the Science and Technology Innovation Committee Foundation of Shenzhen (Grant No. ZDSYS201703031748284).

References

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

[Bartlett *et al.*, 2005] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[Bellet and Habrard, 2015] Aurélien Bellet and Amaury Habrard. Robustness and generalization for metric learning. *Neurocomputing*, 151:259–267, 2015.

[Cao *et al.*, 2016] Qiong Cao, Zheng-Chu Guo, and Yiming Ying. Generalization bounds for metric and similarity learning. *Machine Learning*, 102(1):115–132, 2016.

[Christmann and Zhou, 2016] Andreas Christmann and Ding-Xuan Zhou. On the robustness of regularized pairwise learning methods based on kernels. *Journal of Complexity*, 37:1–33, 2016.

[Clémençon *et al.*, 2008] Stéphan Clémençon, Gabor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *Annals of Statistics*, 36(2):844–874, 2008.

[Cortes and Mohri, 2004] Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. In *NIPS*, pages 313–320, 2004.

[Cucker and Zhou, 2007] Felipe Cucker and Ding-Xuan Zhou. *Learning theory: an approximation theory viewpoint*. Cambridge Univ. Press, Cambridge, 2007.

[Gao and Zhou, 2015] Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In *IJCAI*, pages 939–945, 2015.

[Gao *et al.*, 2013] Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass AUC optimization. In *ICML*, pages 906–914, 2013.

[Guo *et al.*, 2016] Zheng-Chu Guo, Yiming Ying, and Ding-Xuan Zhou. Online regularized learning with pairwise loss functions. *Advances in Computational Mathematics*, pages 1–24, 2016.

[Hu *et al.*, 2015] Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13(04):437–455, 2015.

[Jin *et al.*, 2009] Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning: theory and algorithm. In *NIPS*, pages 862–870, 2009.

[Kakade *et al.*, 2012] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13:1865–1890, 2012.

[Kar *et al.*, 2013] Purushottam Kar, Bharath Sriperumbudur, Prateek Jain, and Harish Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In *ICML*, pages 441–449, 2013.

[Kriukova *et al.*, 2016] Galyna Kriukova, Oleksandra Panasiuk, Sergei V Pereverzyev, and Pavlo Tkachenko. A linear functional strategy for regularized ranking. *Neural Networks*, 73:26–35, 2016.

[Lei and Ying, 2016] Yunwen Lei and Yiming Ying. Generalization analysis of multi-modal metric learning. *Analysis and Applications*, 14(04):503–521, 2016.

[Lin and Zhou, 2018] Shao-Bo Lin and Ding-Xuan Zhou. Distributed kernel-based gradient descent algorithms. *Constructive Approximation*, 47(2):249–276, 2018.

[Lin *et al.*, 2017] Junhong Lin, Yunwen Lei, Bo Zhang, and Ding-Xuan Zhou. Online pairwise learning algorithms with convex loss functions. *Information Sciences*, 406:57–70, 2017.

[Mohri *et al.*, 2012] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2012.

[Mukherjee and Zhou, 2006] Sayan Mukherjee and Ding-Xuan Zhou. Learning coordinate covariances via gradients. *Journal of Machine Learning Research*, 7:519–549, 2006.

[Rejchel, 2012] Wojciech Rejchel. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13(1):1373–1392, 2012.

[Smale and Zhou, 2003] Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(01):17–41, 2003.

[Sridharan *et al.*, 2009] Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *NIPS*, pages 1545–1552, 2009.

[Wang *et al.*, 2012] Yuyang Wang, Roni Khardon, Dmitry Pechyony, and Rosie Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *COLT*, pages 1–22, 2012.

[Xing *et al.*, 2003] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, pages 505–512, 2003.

[Ying and Zhou, 2016] Yiming Ying and Ding-Xuan Zhou. Online pairwise learning algorithms. *Neural Computation*, 28(4):743–777, 2016.

[Ying *et al.*, 2009] Yiming Ying, Kaizhu Huang, and Colin Campbell. Sparse metric learning via smooth optimization. In *NIPS*, pages 2214–2222, 2009.

[Zhang *et al.*, 2017] Lijun Zhang, Tianbao Yang, and Rong Jin. Empirical risk minimization for stochastic convex optimization: $O(1/n)$-and $O(1/n^2)$-type of risk bounds. In *COLT*, pages 1954–1979, 2017.

[Zhao *et al.*, 2011] Peilin Zhao, Rong Jin, Tianbao Yang, and Steven C Hoi. Online AUC maximization. In *ICML*, pages 233–240, 2011.

[Zhao *et al.*, 2017] Yulong Zhao, Jun Fan, and Lei Shi. Learning rates for regularized least squares ranking algorithm. *Analysis and Applications*, 15(06):815–836, 2017.