A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization

Robin Vogel [1 2] **Aurélien Bellet** [3] **Stéphan Clémençon** [1]

Abstract

The performance of many machine learning techniques depends on the choice of an appropriate similarity or distance measure on the input space. Similarity learning (or metric learning) aims at building such a measure from training data so that observations with the same (resp. different) label are as close (resp. far) as possible. In this paper, similarity learning is investigated from the perspective of pairwise bipartite ranking, where the goal is to rank the elements of a database by decreasing order of the probability that they share the same label with some query data point, based on the similarity scores. A natural performance criterion in this setting is pointwise ROC optimization: maximize the true positive rate under a fixed false positive rate. We study this novel perspective on similarity learning through a rigorous probabilistic framework. The empirical version of the problem gives rise to a constrained optimization formulation involving U-statistics, for which we derive universal learning rates as well as faster rates under a noise assumption on the data distribution. We also address the large-scale setting by analyzing the effect of sampling-based approximations. Our theoretical results are supported by illustrative numerical experiments.

1. Introduction

Similarity (or distance) functions play a key role in many machine learning algorithms for problems ranging from classification (e.g., k-nearest neighbors) and clustering (e.g., k-means) to dimensionality reduction (van der Maaten & Hinton, 2008) and ranking (Chechik et al., 2010). The success of such methods are heavily dependent on the relevance of the similarity function to the task and dataset of interest. This has motivated the research in similarity and distance metric learning (Bellet et al., 2015), a line of work which consists in automatically learning a similarity function from data. This training data often comes in the form of pairwise similarity judgments derived from labels, such as positive (resp. negative) pairs composed of two instances with same (resp. different) label. Most existing algorithms can then be framed as unconstrained optimization problems where the objective is to minimize some average loss function over the set of similarity judgments (see for instance Goldberger et al., 2004; Weinberger & Saul, 2009; Bellet et al., 2012, for methods tailored to classification). Some generalization bounds for this class of methods have been derived, accounting for the specific dependence structure found in the training similarity judgments (Jin et al., 2009; Bellet & Habrard, 2015; Cao et al., 2016; Jain et al., 2017; Verma & Branson, 2015). We refer to Kulis (2012) and Bellet et al. (2015) for detailed surveys on similarity and metric learning.

In this paper, we study similarity learning from the perspective of *pairwise bipartite ranking*, where the goal is to rank the elements of a database by decreasing order of the probability that they share the same label with some query data point. This problem is motivated by many concrete applications: for instance, biometric identification aims to check the claimed identity of an individual by matching her biometric information (e.g., a photo taken at an airport) with a large reference database of authorized people (e.g., of passport photos) (Jain et al., 2011). Given a similarity function and a threshold, the database elements are ranked in decreasing order of similarity score with the query, and the matching elements are those whose score is above the threshold. In this context, performance criteria are related to the ROC curve associated with the similarity function, i.e., the relation between the false positive rate and the true positive rate. Previous approaches have empirically tried to optimize the Area under the ROC curve (AUC) of the similarity function (McFee & Lanckriet, 2010; Huo et al., 2018), without establishing any generalization guarantees. The AUC is a global summary of the ROC curve which penalizes pairwise ranking errors regardless of the positions in the list. More local versions of the AUC (e.g., focusing

[1]Télécom ParisTech, Paris, France [2]IDEMIA, Colombes, France [3]INRIA, France. Correspondence to: Robin Vogel <robin.vogel@telecom-paristech.fr>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

on the top of the list) are difficult to optimize in practice and lead to complex nonconvex formulations (Clémençon & Vayatis, 2007; Huo et al., 2018). In contrast, the performance criterion we consider in this work is *pointwise* ROC *optimization*, which aims at maximizing the true positive rate under a fixed false positive rate. This objective, formulated as a constrained optimization problem, naturally expresses the operational constraints present in many practical scenarios. For instance, in biometric applications such as the one outlined above, the verification system is typically set to keep the proportion of people falsely considered a match below a predefined acceptable threshold (see e.g., Jain et al., 2000; 2004).

In addition to proposing an appropriate probabilistic framework to study this novel perspective on similarity learning, we make the following key contributions:

Universal and fast learning rates. We derive statistical guarantees for the approach of solving the constrained optimization problem corresponding to the empirical version of our theoretical objective, based on a dataset of n labeled data points. As the empirical quantities involved are not i.i.d. averages but rather in the form of Ustatistics (Lee, 1990), our results rely on concentration bounds developed for U-processes (Clémençon et al., 2008). We first derive a learning rate of order $O(1/\sqrt{n})$ which holds without any assumption on the data distribution. We then show that one can obtain faster rates under a low-noise assumption on the data distribution, which has the form of a margin criterion involving the conditional quantile. We are unaware of previous results of this kind for constrained similarity/distance metric learning. Interestingly, we are able to illustrate the faster rates empirically through numerical simulations, which is rarely found in the literature on fast learning rates.

Scalability by sampling. We address scalability issues that arise from the very large number of negative pairs when the dataset and the number of classes are large. In particular, we show that using an approximation of the pairwise negative risk consisting of $O(n)$ randomly sampled terms, known as an incomplete U-statistic (see Blom, 1976; Lee, 1990), is sufficient to maintain the universal learning rate of $O(1/\sqrt{n})$. We analyze two different choices of sampling strategies and discuss properties of the data distribution which can make one more accurate than the other. We further provide numerical experiments to illustrate the practical benefits of this strategy.

The rest of this paper is organized as follows. Section 2 introduces the proposed probabilistic framework for similarity learning and draws connections to existing approaches. In Section 3, we derive universal and fast learning rates for the minimizer of the empirical version of our problem. Section 4 addresses scalability issues through random sampling, and Section 5 presents some numerical experiments. Detailed proofs can be found in the supplementary material.

2. Background and Preliminaries

In this section, we introduce the main notations and concepts involved in the subsequent analysis. We formulate the supervised similarity learning problem from the perspective of pairwise bipartite ranking, and highlight connections with some popular metric and similarity learning algorithms of the literature. Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the Dirac mass at any point x by δ_x, and the pseudo-inverse of any cdf $F(u)$ on $\mathbb{R}$ by $F^{-1}(t) = \inf\{v \in \mathbb{R} : \ F(v) \geq t\}$.

2.1. Probabilistic Framework for Similarity Learning

We consider the (multi-class) classification setting. The random variable Y denotes the output label with values in the discrete set $\{1, \ \ldots, \ K\}$ with $K \geq 1$, and X is the input random variable, taking its values in a feature space $\mathcal{X} \subset \mathbb{R}^d$ with $d \geq 1$ and modeling some information hopefully useful to predict Y. We denote by $\mu(dx)$ the marginal distribution of X and by $\eta(x) = (\eta_1(x), \ \ldots, \ \eta_K(x))$ the posterior probability, where $\eta_k(x) = \mathbb{P}\{Y = k \mid X = x\}$ for $x \in \mathcal{X}$ and $k \in \{1, \ \ldots, \ K\}$. The distribution of the random pair (X, Y) is entirely characterized by $P = (\mu, \eta)$. The probability of occurrence of an observation with label $k \in \{1, \ \ldots, \ K\}$ is assumed to be strictly positive and denoted by $p_k = \mathbb{P}\{Y = k\}$, and the conditional distribution of X given $Y = k$ is denoted by $\mu_k(dx)$. Equipped with these notations, we have $\mu = \sum_{k=1}^{K} p_k \mu_k$.

Optimal similarity measures. The objective of *similarity learning* can be informally formulated as follows: the goal is to learn, from a training sample $\mathcal{D}_n = \{(X_1, Y_1), \ \ldots, \ (X_n, Y_n)\}$ composed of $n \geq 1$ independent copies of (X, Y), a (measurable) similarity measure $S : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that given two independent pairs (X, Y) and (X', Y') drawn from P, the larger the similarity $S(X, X')$ between two observations, the more likely they are to share the same label. The set of all similarity measures is denoted by $\mathcal{S}$. The class $\mathcal{S}^*$ of optimal similarity rules naturally corresponds to the set of strictly increasing transforms T of the pairwise posterior probability $\eta(x, x') = \mathbb{P}\{Y = Y' \mid (X, X') = (x, x')\}$, where (X', Y') denotes an independent copy of (X, Y):

$$\{T \circ \eta \,|\, T : Im(\eta) \to \mathbb{R}_+ \text{ borelian, strictly increasing}\},$$

and where $Im(\eta)$ denotes the support of $\eta(X, X')$'s distribution. With the notations previously introduced, we have $\eta(x, x') = \sum_{k=1}^{K} \eta_k(x)\eta_k(x')$ for all $(x, x') \in \mathcal{X}^2$. A similarity rule $S^* \in \mathcal{S}^*$ defines the optimal preorder[1] $\preceq^*$ on

[1] A preorder on a set $\mathcal{X}$ is any reflexive and transitive binary relationship on $\mathcal{X}$. A preorder is an order if, in addition, it is

the product space $\mathcal{X} \times \mathcal{X}$: for any $(x_1, x_2, x_3, x_4) \in \mathcal{X}^4$, x_1 and x_2 are more similar to each other than x_3 and x_4 iff $\eta(x_1, x_2) \geq \eta(x_3, x_4)$, and one writes $(x_3, x_4) \preceq^* (x_1, x_2)$ in this case. For any $x \in \mathcal{X}$, S^* also defines a preorder $\preceq^*_x$ on the input space $\mathcal{X}$, permitting to rank optimally all possible observations by increasing degree of similarity to x: for all $(x_1, x_2) \in \mathcal{X}^2$, x_1 is more similar to x than x_2 (one writes $x_2 \preceq^*_x x_1$) iff $(x, x_2) \preceq^* (x, x_1)$, meaning that $\eta(x, x_2) \leq \eta(x, x_1)$. We point out that, despite its simplicity, this framework covers a wide variety of applications, such as the biometric identification problem mentioned earlier in the introduction.

Similarity learning as pairwise bipartite ranking. In view of the objective formulated above, similarity learning can be seen as a *bipartite ranking* problem on the product space $\mathcal{X} \times \mathcal{X}$ where, given two independent realizations (X, Y) and (X', Y') of P, the input r.v. is the pair (X, X') and the binary label is $Z = 2\mathbb{I}\{Y = Y'\} - 1$. One may refer to *e.g.* Clémençon & Vayatis (2009) and the references therein for a statistical learning view of bipartite ranking. ROC analysis is the gold standard to evaluate the performance of a similarity measure S in this context, *i.e.* to measure how close the preorder induced by S is to $\preceq^*$. The ROC curve of S is the PP-plot $t \in \mathbb{R}_+ \mapsto (F_{S,-}(t), F_{S,+}(t))$, where, for all $t \geq 0$,

$$\begin{aligned} F_{S,-}(t) &= \mathbb{P}\{S(X, X') > t \mid Z = -1\}, \\ F_{S,+}(t) &= \mathbb{P}\{S(X, X') > t \mid Z = +1\}, \end{aligned}$$

where possible jumps are connected by line segments. Hence, it can be viewed as the graph of a continuous function $\alpha \in (0, 1) \mapsto \mathrm{ROC}_S(\alpha)$, where $\mathrm{ROC}_S(\alpha) = F_{S,+} \circ F^{-1}_{S,-}(\alpha)$ at any point $\alpha \in (0, 1)$ such that $F_{S,-} \circ F^{-1}_{S,-}(\alpha) = \alpha$. The curve ROC_S reflects the ability of S to discriminate between pairs with same labels and pairs with different labels: the stochastically smaller than $F_{S,-}$ the distribution $F_{S,+}$ is, the higher the associated ROC curve. Note that it corresponds to the type I error vs power plot of the statistical test $\mathbb{I}\{S(X, X') > t\}$ when the null hypothesis stipulates that X and X' have different marginal distribution (*i.e.*, $Y \neq Y'$). A similarity measure S_1 is said to be more accurate than another similarity S_2 when $\mathrm{ROC}_{S_2}(\alpha) \leq \mathrm{ROC}_{S_1}(\alpha)$ for any $\alpha \in (0, 1)$. A straightforward Neyman-Pearson argument shows that $\mathcal{S}^*$ is the set of optimal elements regarding this partial order on $\mathcal{S}$: $\forall (S, S^*) \in \mathcal{S} \times \mathcal{S}^*, \mathrm{ROC}_S(\alpha) \leq \mathrm{ROC}_{S^*}(\alpha) = \mathrm{ROC}_\eta(\alpha)$ for all $\alpha \in (0, 1)$. For simplicity, we will assume that the conditional cdf of $\eta(X, X')$ given $Z = -1$ is invertible.

Pointwise ROC optimization. In many applications, one is interested in finding a similarity function which optimizes the ROC curve at a particular point $\alpha \in (0, 1)$. The superlevel sets of similarity functions in $\mathcal{S}^*$ define the solutions of pointwise ROC optimization problems in this context. In the above framework, it indeed follows from Neyman Pearson's lemma that the test statistic of type I error less than α with maximum power is the indicator function of the set $\mathcal{R}^*_\alpha = \{(x, x') \in \mathcal{X}^2 : \ \eta(x, x') \geq Q^*_\alpha\}$, where Q^*_α is the conditional quantile of the r.v. $\eta(X, X')$ given $Z = -1$ at level $1 - \alpha$. Restricting our attention to similarity functions bounded by 1, this corresponds to the unique solution of the following problem:

$$\max_{S:\mathcal{X}^2 \to [0,1],\ \text{borelian}} R^+(S) \quad \text{subject to} \quad R^-(S) \leq \alpha, \quad (1)$$

where $R^+(S) = \mathbb{E}[S(X, X') \mid Z = +1]$ is referred to as *positive risk* and $R^-(S) = \mathbb{E}[S(X, X') \mid Z = -1]$ as the *negative risk*.

Remark 1. (UNCONSTRAINED FORMULATION) *The superlevel set $\mathcal{R}^*_\alpha$ of the pairwise posterior probability $\eta(x, x')$ is the measurable subset $\mathcal{R}$ of $\mathcal{X}^2$ that minimizes the cost-sensitive classification risk:*

$$\begin{aligned} p(1 - Q^*_\alpha)\mathbb{P}\left\{(X, X') \notin \mathcal{R} \mid Z = +1\right\} + \\ (1 - p)Q^*_\alpha \mathbb{P}\left\{(X, X') \in \mathcal{R} \mid Z = -1\right\}, \end{aligned}$$

where $p = \mathbb{P}\{Z = +1\} = \sum_{k=1}^K p_k^2$. Notice however that the asymmetry factor, namely the quantile Q^*_α, is unknown in practice, just like the r.v. $\eta(X, X')$. For this reason, one typically considers the problem of maximizing

$$R^+(S) - \lambda R^-(S), \quad (2)$$

for different values of the constant $\lambda > 0$. The performance in terms of ROC *curve can only be analyzed* a posteriori*, and the value λ thus needs to be tuned empirically by model selection techniques.*

2.2. Connections to Existing Similarity and Metric Learning Approaches

We point out that the similarity learning framework described above can be equivalently described in terms of learning a dissimilarity measure (or pseudo distance metric) $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. In this case, the pointwise ROC optimization problem (1) translates into:

$$\begin{aligned} \min_{D:\mathcal{X}^2 \to [0,1]} \mathbb{E}\left[D(X, X') \mid Z = +1\right] \\ \text{subject to } \mathbb{E}\left[D(X, X') \mid Z = -1\right] \geq 1 - \alpha. \end{aligned} \quad (3)$$

A large variety of practical similarity and distance metric learning algorithms have been proposed in the literature, all revolving around the same idea that a good similarity function should output large scores for pairs of points in the same class, and small scores for pairs with different label. They

antisymmetrical.

differ from one another by the class of metric/similarity functions considered, and by the kind of objective function they optimize (see Bellet et al., 2015, for a comprehensive review). In any case, ROC curves are commonly used to evaluate metric learning algorithms when the number of classes is large (see for instance Guillaumin et al., 2009; Kstinger et al., 2012; Shen et al., 2012), which makes our framework very relevant in practice. Several popular algorithms optimize an empirical version of Problems (1)-(3), often in their unconstrained version as in (2) (Liu et al., 2010; Xie & Xing, 2015). We argue here in favor of the constrained version as the parameter α has a direct correspondence with the point $\mathrm{ROC}_S(\alpha)$ of the ROC curve, unlike the unconstrained case (see Remark 1). This will be illustrated in our numerical experiments of Section 5.

Interestingly, our framework sheds light on MMC, the seminal metric learning algorithm of Xing et al. (2002) originally designed for clustering with side information. MMC solves the empirical version of (3) with α fixed to 0. This is because MMC optimizes over a class of distance functions with unbounded values, hence modifying α does not change the solution (up to a scaling factor). We note that by choosing a bounded family of distance functions, one can use the same formulation to optimize the pointwise ROC curve.

3. Statistical Guarantees for Generalization

Pointwise ROC optimization problems have been investigated from a statistical learning perspective by Scott & Nowak (2005) and Clémençon & Vayatis (2010) in the context of binary classification. The major difference with the present framework lies in the pairwise nature of the quantities appearing in Problem (1) and, consequently, in the complexity of its empirical version. In particular, natural statistical estimates for the positive risk $R^+(S)$ and the negative risk $R^-(S)$ (1) computed on the training sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are given by:

$$\widehat{R}_n^+(S) = \frac{1}{n_+} \sum_{1 \leq i < j \leq n} S(X_i, X_j) \cdot \mathbb{I}\{Y_i = Y_j\}, \quad (4)$$

$$\widehat{R}_n^-(S) = \frac{1}{n_-} \sum_{1 \leq i < j \leq n} S(X_i, X_j) \cdot \mathbb{I}\{Y_i \neq Y_j\}, \quad (5)$$

where $n_+ = \sum_{1 \leq i < j \leq n} \mathbb{I}\{Y_i = Y_j\} = n(n-1)/2 - n_-$. It is important to note that these quantities are not i.i.d. averages, since several pairs involve each i.i.d. sample. This breaks the analysis carried out by Clémençon & Vayatis (2010, Section 5 therein) for the case of binary classification.

We can however observe that $U_n^+(S) = 2n_+/(n(n-1))\widehat{R}_n^+(S)$ and $U_n^-(S) = 2n_-/(n(n-1))\widehat{R}_n^-(S)$ are U-statistics of degree two with respective symmetric kernels $h_+((x,y),(x',y')) = S(x,x') \cdot \mathbb{I}\{y = y'\}$ and $h_-((x,y),(x',y')) = S(x,x') \cdot \mathbb{I}\{y \neq y'\}$.[2] We will therefore be able to use existing representation tricks to derive concentration bounds for U-processes (collections of U-statistics indexed by classes of kernel functions), under appropriate complexity conditions, see *e.g.* (Dudley, 1999).

We thus investigate the generalization ability of solutions obtained by solving the empirical version of Problem (1), where we also restrict the domain to a subset $\mathcal{S}_0 \subset \mathcal{S}$ of similarity functions bounded by 1, and we assume $\mathcal{S}_0$ has controlled complexity (*e.g.* finite VC dimension). Finally, we replace the target level α by $\alpha + \Phi$, where Φ is some tolerance parameter that should be of the same order as the maximal deviation $\sup_{S \in \mathcal{S}_0} |\widehat{R}_n^-(S) - R^-(S)|$. This leads to the following empirical problem:

$$\max_{S \in \mathcal{S}_0} \widehat{R}_n^+(S) \quad \text{subject to} \quad \widehat{R}_n^-(S) \leq \alpha + \Phi. \quad (6)$$

Following Clémençon et al. (2008), we have the following lemma.

Lemma 1. *(Clémençon et al., 2008, Corollary 3) Assume that $\mathcal{S}_0$ is a VC-major class of functions with finite* VC *dimension $V < +\infty$. We have with probability larger than $1 - \delta$: $\forall n > 1$,*

$$\sup_{S \in \mathcal{S}_0} \left| \widehat{U}_n^+(S) - \mathbb{E}[\widehat{U}_n^+(S)] \right| \leq 2C\sqrt{\frac{V}{n}} + 2\sqrt{\frac{\log(1/\delta)}{n-1}}, \quad (7)$$

where C is a universal constant, explicited in Bousquet et al. (2004, page 198 therein).

A similar result holds for the U-process $\{\widehat{U}_n^-(S) - U^-(S)\}_{S \in \mathcal{S}_0}$. We are now ready to state our universal learning rate, describing the generalization capacity of solutions of the constrained optimization program (6) under specific conditions for the class $\mathcal{S}_0$ of similarity functions and a suitable choice of the tolerance parameter Φ. This result can be established by combining Lemma 1 with the derivations of Clémençon & Vayatis (2010, Theorem 10 therein). Details can be found in the supplementary material.

Theorem 1. *Suppose that the assumptions of Lemma 1 are fulfilled and that $S(x,x') \leq 1$ for all $S \in \mathcal{S}_0$ and any $(x,x') \in \mathcal{X}^2$. Assume also that there exists a constant $\kappa \in (0,1)$ such that $\kappa \leq \sum_{k=1} p_k^2 \leq 1 - \kappa$. For all $\delta \in (0,1)$ and $n > 1$, set*

$$\Phi_{n,\delta} = 2C\kappa^{-1}\sqrt{\frac{V}{n}} + 2\kappa^{-1}(1+\kappa^{-1})\sqrt{\frac{\log(3/\delta)}{n-1}},$$

and consider a solution $\hat{S}_n$ of the contrained minimization problem (6) *with $\Phi = \Phi_{n,\delta/2}$. Then, for any $\delta \in (0,1)$,*

[2] We give the definition of U-statistics in the supplementary material for completeness.

we have simultaneously with probability at least $1-\delta$*:* $\forall n \geq 1 + 4\kappa^{-2}\log(3/\delta)$*,*

$$R^+(\hat{S}_n) \geq \mathrm{ROC}_{S^*}(\alpha) - \Phi_{n,\delta/2} - \Big\{\mathrm{ROC}_{S^*}(\alpha) - \sup_{S\in\mathcal{S}_0:\, R^-(S)\leq\alpha} R^+(S)\Big\}, \quad (8)$$

and

$$R^-(\hat{S}_n) \leq \alpha + \Phi_{n,\delta/2}. \quad (9)$$

Remark 2. (ON BIAS AND MODEL SELECTION) *We point out that the last term on the right hand side of* (8) *should be interpreted as the bias of the statistical learning problem* (6)*, which depends on the richness of class* $\mathcal{S}_0$*. This term vanishes when* $\mathbb{I}\{(x,x') \in \mathcal{R}^*_\alpha\}$ *belongs to* $\mathcal{S}_0$*. Choosing a class yielding a similarity rule of highest true positive rate with large probability can be tackled by means of classical model selection techniques, based on resampling methods or complexity penalization (note that oracle inequalities can be straightforwardly derived from the same analysis).*

Except for the minor condition stipulating that the probability of occurrence of "positive pairs" $\sum_{k=1}^K p_k^2$ stays bounded away from 0 and 1, the generalization bound stated in Theorem 1 holds whatever the probability distribution of (X,Y). Beyond such universal results, we investigate situations where rates faster than $O(1/\sqrt{n})$ can be achieved by solutions of (6). Such fast rates results exist for binary classification under the so-called Mammen-Tsybakov noise condition, see *e.g.* Bousquet et al. (2004) for details. By means of a variant of the Bernstein inequality for U-statistics, we can establish fast rate bounds under the following condition on the data distribution.

Noise assumption (NA). *There exist a constant* c *and* $a \in [0,1]$ *such that, almost surely,*

$$\mathbb{E}_{X'}\big[|\eta(X,X') - Q^*_\alpha|^{-a}\big] \leq c.$$

This noise condition is similar to that introduced by Mammen & Tsybakov (1995) for the binary classification framework, except that the threshold $1/2$ is replaced here by the conditional quantile Q^*_α. It characterizes "nice" distributions for the problem of ROC optimization at point α: it essentially ensures that the pairwise posterior probability is bounded away from Q^*_α with high probability. Under the assumption, we can derive the following fast learning rates.

Theorem 2. *Suppose that the assumptions of Theorem 1 are satisfied, that condition NA holds true and that the optimal similarity rule* $S^*_\alpha(x,x') = \mathbb{I}\{(x,x') \in \mathcal{R}^*_\alpha\}$ *belongs to* $\mathcal{S}_0$*. Fix* $\delta > 0$*. Then, there exists a constant* C'*, depending on* δ*,* κ*,* Q^*_α*,* a*,* c *and* V *such that, with probability at least* $1-\delta$*,*

$$\mathrm{ROC}_{S^*}(\alpha) - R^+(\hat{S}_n) \leq C' n^{-(2+a)/4}, \quad \textit{and} \quad R^-(\hat{S}_n) \leq \alpha + 2\Phi_{n,\delta/2}.$$

Remark 3. (ON THE NA CONDITION) *The noise condition is automatically fulfilled for any* $a \in (0,1)$ *when, for almost every point* x *with respect to the measure induced by* X*,* $\eta(x,X')$ *has an absolutely continuous distribution and bounded density. This assumption means that the problem of ranking by similarity to an instance* x *is not too hard for any value of* x*, see supplementary material for more details.*

The proof is based on the same argument as that of Clémençon & Vayatis (2010, Theorem 12 therein), except that it involves a sharp control of the fluctuations of the U-statistic estimates of the true positive rate excess $\mathrm{ROC}_{S^*}(\alpha) - R^+(S)$ over the class $\mathcal{S}_0$. The reduced variance property of U-statistics plays a crucial role in the analysis, which essentially relies on the Hoeffding decomposition (see Hoeffding, 1948). Technical details can be found in the supplementary material.

4. Scalability by Sampling Approximations

In the previous section, we analyzed the learning rates achieved by a minimizer of the empirical problem (6). In the large-scale setting, solving this problem can be computationally costly due to the very large number of training pairs. In particular, the positive and negative empirical risks $\widehat{R}^+_n(S)$ and $\widehat{R}^-_n(S)$ are sums over respectively $\sum_{k=1}^K n_k(n_k-1)/2$ and $\sum_{k<l} n_k n_l$ pairs. We focus here more specifically on the setting where we have a large number of (rather balanced) classes, as in our biometric identification motivating example where a class corresponds to an identity. In this regime, we are facing a highly imbalanced problem since the number of negative pairs becomes overwhelmingly large compared to the number of positive pairs. For instance, even for the MNIST dataset where the number of classes is only $K = 10$ and $n_k = 6000$, there are already 10 times more negative pairs than positive pairs.

A natural strategy, often used by metric learning practitioners (see e.g., Babenko et al., 2009; Wu et al., 2013; Xie & Xing, 2015), is to drastically subsample the negative pairs while keeping all positive pairs. In this section, we shed light on this popular practice by analyzing the effect of subsampling (conditionally upon the data) the negative pairs onto the generalization performance.

A simple approach consists in replacing the empirical negative risk $\widehat{R}^-_n(S)$ by the following approximation:

$$\bar{R}^-_B(S) := \frac{1}{B}\sum_{(i,j)\in\mathcal{P}_B} S(X_i, X_j),$$

where $\mathcal{P}_B$ is a set of cardinality B built by sampling with replacement in the set of negative training pairs $\Lambda_{\mathcal{P}} = \{(i,j)\,|\,i,j \in \{1,\ldots,n\}; Y_i \neq Y_j\}$. Conditioned upon the

n_k's, $\bar{R}_B^-(S)$ can be viewed as an *incomplete* version of the U-statistic $\widehat{R}_n^-(S)$ consisting of B pairs (Blom, 1976; Lee, 1990).

Despite the simplicity of the above approximation, we also consider an alternative sampling strategy, which consists in sampling a number B of K-tuples containing one random sample of each class. Formally, this corresponds to the following approximation:

$$\widetilde{R}_B^-(S) := \frac{1}{B} \sum_{(i_1,\dots i_K)\in\mathcal{T}_B} h_S(X_{i_1},\dots,X_{i_K}),$$

where $h_S(X_1,\dots,X_K) = \frac{1}{n_-}\sum_{k<l} n_k n_l S(X_k,X_l)$ and $\mathcal{T}_B$ is a set of cardinality B built by sampling with replacement in the set of K-tuples $\Lambda_{\mathcal{T}} = \{(i_1,\dots,i_K)\,|\,i_k\in\{1,\dots,n_k\}; k=1,\dots,K\}$. $\widetilde{R}_B^-(S)$ is also an incomplete version of $\widehat{R}_n^-(S)$, with the alternative view of $\widehat{R}_n^-(S)$ as a *generalized* K-sample U-statistic (Lee, 1990) of degree $(1,\dots,1)$ and kernel h_S, see supplementary material for a full definition. Note that $\widetilde{R}_B^-(S)$ contains $BK(K-1)/2$ pairs, balanced across all class pairs.

$\bar{R}_B^-(S)$ and $\widetilde{R}_B^-(S)$ are both unbiased estimates of $\widehat{R}_n^-(S)$, but their variances are different and one approximation might be better than the other in some regimes. The following result provides expressions for the variances of both incomplete estimators for a fixed budget of B_0 sampled pairs, under a standard asymptotic framework.

Proposition 1. *Let B_0 be the number of pairs sampled in both schemes, and denote $V_n = Var(\hat{R}_n^-(S))$. When $B_0/n \to 0$, $n\to\infty$ and for all $k\in\{1,\dots,K\}$, $n_k/n \to p_k > 0$, we have:*

$$Var(\widetilde{R}_B^-(S)) - V_n \sim \frac{K(K-1)}{2B_0} Var(h_S(X^{(1)},\dots,X^{(k)})),$$
$$Var(\bar{R}_B^-(S)) - V_n \sim B_0^{-1} Var(S(X,X')\,|\,Y\neq Y'),$$

where $X^{(k)}$ denotes $X\,|\,Y=k$ for all $k\in\{1,\dots,K\}$.

Proposition 1 states that if the variance of similarity scores on the negative pairs is high compared to the variance of a weighted average of similarity scores on all types (k,l) of negative pairs, then one should prefer tuple-based sampling (otherwise pair-based sampling is better). As an example, consider the case where the similarity scores on the negative pairs constructed from classes (k_0,l_0) are consistently higher than for other negative pairs. These high similarity pairs will not be sampled very often by the pair-based sampling method, in contrast to the tuple-based approach. In that scenario, the variance of $S(X,X')\,|\,Y\neq Y'$ is high while the variance of $h_S\left(X^{(1)},\dots,X^{(k)}\right)$ is low, and the tuple-based method should be preferred. In practice, the properties of the data should guide the choice of the sampling approach.

We now analyze the effect of sampling on the performance of the empirical risk minimizer. We consider tuple-based sampling (results of the same order can be obtained for pair-based sampling). Let $\tilde{S}_B$ be the minimizer of the following simpler empirical problem:

$$\arg\max_{S\in\mathcal{S}_0} \widehat{R}_n^+(S) \quad \text{subject to } \widetilde{R}_B^-(S) \leq \alpha + \Phi_{n,\delta,B}. \quad (10)$$

We have the following theorem, based on combining Theorem 1 with a result bounding the maximal deviation between $\widehat{R}_n^-(S)$ and its incomplete version $\widetilde{R}_B^-(S)$, see Clémençon et al. (2016).

Theorem 3. *Let $N = \min_{1\leq k\leq K} n_k$ and $\alpha\in(0,1)$, assume that $S^*\in\mathcal{S}_0$ and that $\mathcal{S}_0$ is a VC-major class of dimension V. For all $(\delta,n,B)\in(0,1)\times\mathbb{N}^*\times\mathbb{N}^*$, set*

$$\Phi_{n,\delta,B} = 4\sqrt{\frac{V\log(1+N)}{N}} + \sqrt{\frac{\log(2/\delta)}{N}} + \sqrt{2\frac{V\log(1+\prod_{k=1}^K n_k) + \log(4/\delta)}{B}}.$$

Then we have simultaneously with probability at least $1-\delta$,

$$R^+(\tilde{S}_B) \geq R_*^+ - 2\Phi_{n,\delta,B} \quad \text{and} \quad R^-(\tilde{S}_B) \leq \alpha + 2\Phi_{n,\delta,B}.$$

This result is very similar to Theorem 1, with an additive error term in $O(\sqrt{\log n/B})$. Remarkably, this implies that it is sufficient to sample $B = O(n)$ tuples (hence only $O(nK^2)$ pairs) to preserve the $O(\sqrt{\log n/n})$ learning rate achieved when using all negative pairs. This will be confirmed empirically in our numerical experiments.

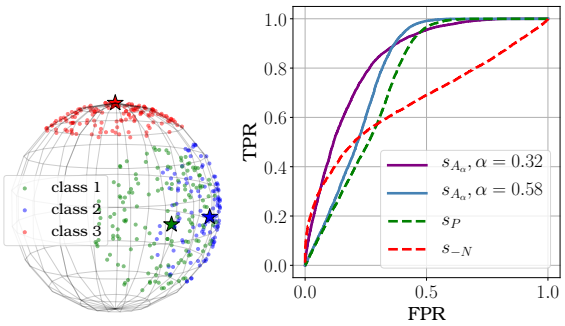
Remark 4 (Approximating the positive risk)**.** *When needed, sampling-based techniques can also be used to approximate the empirical positive risk $\widehat{R}_n^+(S)$, with generalization results analogous to Theorem 3. Details are left to the reader.*

5. Illustrative Experiments

In this section, we present some experiments to illustrate our main results. We first illustrate how solving instances of Problem (6) allows to optimize for specific points of the ROC curve. We then provide some numerical evidence of the fast rates of Theorem 2. Finally, we illustrate our scalability results of Section 4 by showing that dramatically subsampling the negative empirical risk leads to negligible loss in generalization performance.

5.1. Pointwise ROC Optimization

We illustrate on synthetic data that solving (6) for different values of α can optimize for different regions of the



(a) Simulated data (stars are class centroids)

(b) ROC curves

Figure 1. Illustrative experiments for pointwise ROC optimization.

ROC curve. Let $\mathcal{X} \subset \mathbb{R}^d$, and let $\mathcal{S}_0$ be the set of bilinear similarities with norm-constrained matrices

$$\mathcal{S}_0 = \left\{ S_A : (x, x') \mapsto \frac{1}{2}\left(1 + x^\top A x'\right) \;\middle|\; \|A\|_F^2 \leq 1 \right\},$$

where $\|A\|_F^2 = \sum_{i,j=1}^d a_{ij}^2$. Note that when data is scaled ($\|x\| = 1$ for all $x \in \mathcal{X}$), we have $S_A(x, x') \in [0, 1]$ for all $x, x' \in \mathcal{X}$ and all $S_A \in \mathcal{S}_0$. In our simple experiment, we have $K = 3$ classes and observations belong to the sphere in $\mathbb{R}^3$. Denoting by θ_{x,c_i} the angle between the element x and the centroid c_i of class i, we set for all $i \in \{1, 2, 3\}$,

$$\mu_i(x) \propto \mathbb{I}\left\{\theta_{x,c_i} < \frac{\pi}{4}\right\}, \quad p_i = \frac{1}{3}$$

and $c_1 = (\cos(\pi/3), \sin(\pi/3), 0)$, $c_2 = e_2$, $c_3 = e_3$ with e_i vectors of the standard basis of $\mathbb{R}^3$. See Figure 1(a) for a graphical representation of the data.

The solutions of the problem can be expressed in closed form using Lagrangian duality. In particular, when the constraints are saturated, the solution S_{A_α} is an increasing transformation of $s_{P-\lambda_\alpha N}$ with

$$P = \frac{1}{2n_+} \sum_{1 \leq i < j \leq n} \mathbb{I}\{Y_i = Y_j\} \cdot \left(X_i X_j^\top + X_j X_i^\top\right),$$

$$N = \frac{1}{2n_-} \sum_{1 \leq i < j \leq n} \mathbb{I}\{Y_i \neq Y_j\} \cdot \left(X_i X_j^\top + X_j X_i^\top\right),$$

and λ_α is a positive Lagrange multiplier decreasing in α, see supplementary material for details. By varying α, we trade-off between the information contained in the positive pairs (α large, λ_α close to zero) and in the negative pairs (α small, λ_α large), which indeed results in optimizing different areas of the ROC curve, see Figure 1(b).

5.2. Fast Rates

Theorem 2 shows that when the noise assumption NA is verified, faster rates of generalization can be achieved. Showing

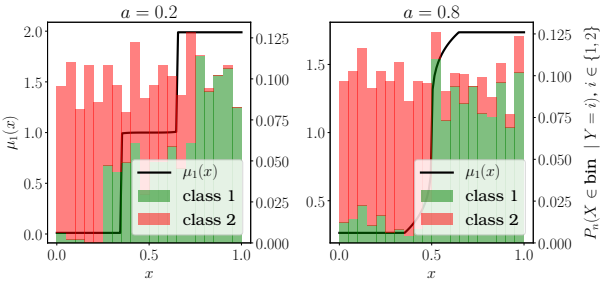


Figure 2. Example distributions and μ_1's for $n = 1000$ and two values of a.

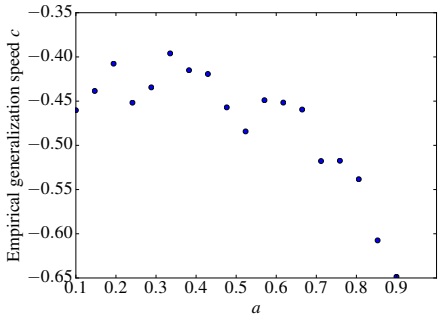


Figure 3. Generalization speed for different values of a.

the existence of fast rates experimentally requires us to design a problem for which the η satisfies NA, which is not trivial due to the pairwise nature of the involved quantities. We emphasize that such empirical evidence of fast rates is rarely found in the literature.

We put ourselves in a simple scenario where $\mathcal{X} = [0, 1]$, $\mu = 1$, $K = 2$ and $p_1 = p_2 = 1/2$. In that context, characterizing $\mu_1(dx)$ is sufficient to have a fully defined problem. With $m \in (0, \frac{1}{2})$, $a \in (0, 1)$ and $C \in (0, \frac{1}{2})$, we set

$$\mu_1(x) = \begin{cases} 2C & \text{if} \quad x \in [0, m], \\ 1 - |2x - 1|^{(1-a)/a} & \text{if} \quad x \in (m, 1/2], \end{cases}$$

where C is chosen so that $Q_\alpha^* = 1/2$ and m is fixed in advance. Since $\int \mu_1(dx) = 1$, we chose μ_1 symmetric in $(1/2, 1)$ to satisfy that constraint. Figure 2 shows example distributions.

Given that $\mu = 1$, the noise assumption with a close to 1 requires that there are sharp variations of η close to Q_α^*. To induce the form of the function more easily, we fixed $Q_\alpha^* = 1/2$, which requires us to choose μ_1 such that the value of the integral of η is controlled while η has the expected local property around $1/2$. More details about the design of the experiment can be found in the supplementary material. When t is small enough, $\mathbb{P}\left(|\eta(X, X') - Q_\alpha^*| \leq t\right)$

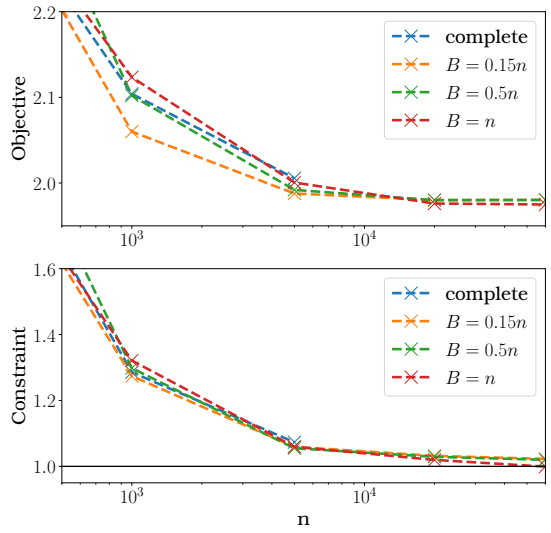


Figure 4. Value of objective and constraint on the test set for various levels of approximation of the negative risk, averaged on 5 runs for each set of parameters (n, B).

is of order $-t^{\frac{a}{1-a}}\log(t)$. Due to the logarithm term in the noise condition, we expect that the generalization speeds to be slightly worse than $O(n^{-(2+a)/4})$.

The family $\mathcal{S}_0$ is composed of indicators of sets, which are parameterized by $t \in (0, 1)$ (see supplementary material for a graphical representation). Each set contains the pairs (x, x') such that one of the supremum distances between (x, x') and $(0, 0)$ or $(1, 1)$ is smaller than t, which writes

$$\{x, x' \in \mathcal{X} \mid \min(\max(1 - x, 1 - x'), \max(x, x')) < t\}\,.$$

The optimal set can thus always be identified, and $R^+(S)$ and $R^-(S)$ can be expressed analytically for some $S \in \mathcal{S}_0$. The empirical problem Eq. (6) is always solved neglecting the tolerance parameter Φ, i.e. setting $\Phi = 0$.

Figure 3 shows experiments for the case $\alpha = 0.26$, $m = 0.35$ and $a \in [0.1, 0.9]$. For some a, the empirical 90-quantile of $\mathrm{ROC}_{S^*}(\alpha) - R^+(\hat{S}_n)$ is computed for different values of n on 1000 experiments and its logarithm is fitted to $C_a \times \log(n) + D_a$ to get the empirical generalization speed C_a. There is a clear downward trend when a increases, illustrating the fast rates in practice.

5.3. Scalability by Sampling

We illustrate the results of Section 4 on MMC (Xing et al., 2002), a popular metric learning algorithm whose formulation is very close to the one we consider. We introduce the set of Mahalanobis distances d_A indexed by a positive semidefinite matrix A:

$$d_A(x, x') = \sqrt{(x - x')^\top A (x - x')}.$$

MMC solves the following problem (using projected gradient ascent):

$$\max_A \frac{1}{n_-} \sum_{1 \le i < j \le n} \mathbb{I}\{Y_i \neq Y_j\} \cdot d_A(X_i, X_j)$$
$$\text{s.t.} \frac{1}{n_+} \sum_{1 \le i < j \le n} \mathbb{I}\{Y_i = Y_j\} \cdot d_A^2(X_i, X_j) \le 1$$
$$A \succeq 0$$

We use MNIST dataset, composed of $70,000$ images representing the 0-9 handwritten digits, with classes roughly equally distributed. We randomly split it into a training set and a test set of $10,000$ instances. As done in previous work, the dimension of the features is reduced using PCA to keep 90% of the explained variance. We approximate the average over negative pairs by sampling K-tuples with B terms, as proposed in Section 4 (pair-based sampling performs similarly on this dataset). We aim to show that optimizing the criterion on the resulting smaller set of pairs does not significantly impact the learning rate (yet greatly reduces training time). We solve MMC on the training set for a varying number of training instances n and of K-tuples B, and report the objective and constraint values on the test set. The results, summarized in Figure 4, confirm the small performance loss due to subsampling, for a huge improvement in terms of computing time. Indeed, when $n = 60,000$, the total number of negative pairs is almost 2 billions while $B = 0.15n$ corresponds to sampling only $400,000$ pairs.

6. Conclusion

We have introduced a rigorous probability framework to study similarity learning from the novel perspective of pairwise bipartite ranking and pointwise ROC optimization. We derived statistical guarantees for generalization in this context, and analyzed the impact of using sampling-based approximations. Our results are illustrated on a series of numerical experiments. Our study opens promising directions of future work. We are especially interested in extending our results to allow the rejection of queries from unseen classes (e.g., unknown identities) at test time (see for instance Bendale & Boult, 2015). This could be achieved by incorporating a loss function to encourage the score of all positive pairs to be above some fixed threshold, below which we would reject the query.

Acknowledgments

This work was supported by IDEMIA. We would like to thank Anne Sabourin for her substantial feedback that has greatly improved this work, as well as the ICML reviewers for their constructive input.

References

Babenko, B., Branson, S., and Belongie, S. J. Similarity metrics for categorization: From monolithic to category specific. In *ICCV*, 2009.

Bellet, A. and Habrard, A. Robustness and Generalization for Metric Learning. *Neurocomputing*, 151(1):259–267, 2015.

Bellet, A., Habrard, A., and Sebban, M. Similarity Learning for Provably Accurate Sparse Linear Classification. In *ICML*, 2012.

Bellet, A., Habrard, A., and Sebban, M. *Metric Learning*. Morgan & Claypool Publishers, 2015.

Bendale, A. and Boult, T. E. Towards Open World Recognition. In *CVPR*, 2015.

Blom, G. Some properties of incomplete U-statistics. *Biometrika*, 63(3):573–580, 1976.

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pp. 169–207. 2004.

Cao, Q., Guo, Z.-C., and Ying, Y. Generalization Bounds for Metric and Similarity Learning. *Machine Learning*, 102(1):115–132, 2016.

Chechik, G., Sharma, V., Shalit, U., and Bengio, S. Large Scale Online Learning of Image Similarity Through Ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010.

Clémençon, S. and Vayatis, N. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

Clémençon, S. and Vayatis, N. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

Clémençon, S., Colin, I., and Bellet, A. Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics. *Journal of Machine Learning Research*, 17 (76):1–36, 2016.

Clémençon, S. and Vayatis, N. Overlaying classifiers: A practical approach to optimal scoring. *Constructive Approximation*, 32(3):619–648, 2010.

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and Empirical Minimization of U-Statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Dudley, R. M. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

Goldberger, J., Roweis, S., Hinton, G., and Salakhutdinov, R. Neighbourhood Components Analysis. In *NIPS*, 2004.

Guillaumin, M., Verbeek, J., and Schmid, C. Is that you? Metric Learning Approaches for Face Identification. In *CVPR*, 2009.

Hoeffding, W. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19:293–325, 1948.

Huo, J., Gao, Y., Shi, Y., and Yin, H. Cross-modal metric learning for auc optimization. *IEEE Transactions on Neural Networks and Learning Systems*, PP(99):1–13, 2018.

Jain, A., Hong, L., and Pankanti, S. Biometric identification. *Communications of the ACM*, 43(2):90–98, 2000.

Jain, A. K., Ross, A., and Prabhakar, S. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.

Jain, A. K., Ross, A. A., and Nandakumar, K. *Introduction to Biometrics*. Springer, 2011.

Jain, L., Mason, B., and Nowak, R. Learning Low-Dimensional Metrics. In *NIPS*, 2017.

Jin, R., Wang, S., and Zhou, Y. Regularized Distance Metric Learning: Theory and Algorithm. In *NIPS*, 2009.

Kulis, B. Metric Learning: A Survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012.

Kstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

Lee, A. J. *U-statistics: Theory and practice*. Marcel Dekker, Inc., New York, 1990.

Liu, W., Tian, X., Tao, D., and Liu, J. Constrained Metric Learning Via Distance Gap Maximization. In *AAAI*, 2010.

Mammen, E. and Tsybakov, A. B. Asympotical minimax recovery of the sets with smooth boundaries. *The Annals of Statistics*, 23(2):502–524, 1995.

McFee, B. and Lanckriet, G. R. G. Metric Learning to Rank. In *ICML*, 2010.

Scott, C. and Nowak, R. A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, Nov 2005.

Shen, C., Kim, J., Wang, L., and van den Hengel, A. Positive Semidefinite Metric Learning Using Boosting-like Algorithms. *Journal of Machine Learning Research*, 13: 1007–1036, 2012.

van der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Verma, N. and Branson, K. Sample complexity of learning mahalanobis distance metrics. In *NIPS*, 2015.

Weinberger, K. Q. and Saul, L. K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

Wu, P., Hoi, S. C., Xia, H., Zhao, P., Wang, D., and Miao, C. Online multimodal deep similarity learning with application to image retrieval. In *ACM Multimedia*, 2013.

Xie, P. and Xing, E. P. Large Scale Distributed Distance Metric Learning. Technical report, arXiv:1412.5949, 2015.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. J. Distance Metric Learning with Application to Clustering with Side-Information. In *NIPS*, 2002.