Sample Complexity of Learning Mahalanobis Distance Metrics

Nakul Verma
Janelia Research Campus, HHMI
verman@janelia.hhmi.org

Kristin Branson
Janelia Research Campus, HHMI
bransonk@janelia.hhmi.org

Abstract

Metric learning seeks a transformation of the feature space that enhances prediction quality for a given task. In this work we provide PAC-style sample complexity rates for supervised metric learning. We give matching lower- and upper-bounds showing that sample complexity scales with the representation dimension when no assumptions are made about the underlying data distribution. In addition, by leveraging the structure of the data distribution, we provide rates *fine-tuned* to a specific notion of the intrinsic complexity of a given dataset, allowing us to relax the dependence on representation dimension. We show both theoretically and empirically that augmenting the metric learning optimization criterion with a simple norm-based regularization is important and can help *adapt* to a dataset's intrinsic complexity yielding better generalization, thus partly explaining the empirical success of similar regularizations reported in previous works.

1 Introduction

In many machine learning tasks, data is represented in a high-dimensional Euclidean space. The L_2 distance in this space is then used to compare observations in methods such as clustering and nearest-neighbor classification. Often, this distance is not ideal for the task at hand. For example, the presence of uninformative or mutually correlated measurements arbitrarily inflates the distances between pairs of observations. *Metric learning* has emerged as a powerful technique to learn a *metric* in the representation space that emphasizes feature combinations that improve prediction while suppressing spurious measurements. This has been done by exploiting class labels [1, 2] or other forms of supervision [3] to find a Mahalanobis distance metric that respects these annotations.

Despite the popularity of metric learning methods, few works have studied how problem complexity scales with key attributes of the dataset. In particular, how do we expect generalization error to scale—both theoretically and practically—as one varies the number of informative and uninformative measurements, or changes the noise levels? In this work, we develop two general frameworks for PAC-style analysis of supervised metric learning. The *distance-based* metric learning framework uses class label information to derive distance constraints. The objective is to learn a metric that yields smaller distances between examples from the same class than those from different classes. Algorithms that optimize such distance-based objectives include Mahalanobis Metric for Clustering (MMC) [4], Large Margin Nearest Neighbor (LMNN) [1] and Information Theoretic Metric Learning (ITML) [2]. Instead of using distance comparisons as a proxy, however, one can also optimize for a specific prediction task directly. The second framework, the *classifier-based* metric learning framework, explicitly incorporates the hypotheses associated with the prediction task to learn effective distance metrics. Examples in this regime include [5] and [6].

Our analysis shows that in both frameworks, the sample complexity scales with a dataset's representation dimension (Theorems 1 and 3), and this dependence is necessary in the absence of assumptions about the underlying data distribution (Theorems 2 and 4). By considering any Lipschitz loss, our results improve upon previous sample complexity results (see Section 6) and, for the first time, provide matching lower bounds.

In light of our observation that data measurements often include uninformative or weakly informative features, we expect a metric that yields good generalization performance to de-emphasize such features and accentuate the relevant ones. We thus formalize the *metric learning complexity* of a given dataset in terms of the intrinsic complexity d of the optimal metric. For Mahalanobis metrics, we characterize intrinsic complexity by the *norm* of the matrix representation of the metric. We refine our sample complexity results and show a *dataset-dependent* bound for both frameworks that relaxes the dependence on representation dimension and instead scales with the dataset's intrinsic metric learning complexity d (Theorem 7).

Based on our dataset-dependent result, we propose a simple variation on the empirical risk minimizing (ERM) algorithm that returns a metric (of complexity d) that jointly minimizes the observed sample bias and the expected intra-class variance for metrics of fixed complexity d. This bias-variance balancing criterion can be viewed as a structural risk minimizing algorithm that provides better generalization performance than an ERM algorithm and justifies norm-regularization of weighting metrics in the optimization criteria for metric learning, partly explaining empirical success of similar objectives [7, 8]. We experimentally validate how the basic principle of norm-regularization can help enhance the prediction quality even for existing metric learning algorithms on benchmark datasets (Section 5). Our experiments highlight that norm-regularization indeed helps learn weighting metrics that better adapt to the signal in data in high-noise regimes.

2 Preliminaries

In this section, we define our notation, and explicitly define the distance-based and classifier-based learning frameworks. Given a D-dimensional representation space $X = \mathbb{R}^D$, we want to learn a weighting, or a *metric*[1] M^* on X that minimizes some notion of *error* on data drawn from a fixed unknown distribution $\mathcal{D}$ on $X \times \{0,1\}$:

$$M^* := \operatorname{argmin}_{M \in \mathcal{M}} \operatorname{err}(M, \mathcal{D}),$$

where $\mathcal{M}$ is the class of weighting metrics $\mathcal{M} := \{M \mid M \in \mathbb{R}^{D \times D}, \sigma_{\max}(M) = 1\}$ (we constrain the maximum singular value $\sigma_{\max}$ to remove arbitrary scalings). For supervised metric learning, this *error* is typically label-based and can be defined in two intuitive ways.

The **distance-based framework** prefers metrics M that bring data from the same class closer together than those from opposite classes. The corresponding distance-based error then measures how the distances amongst data violate class labels:

$$\operatorname{err}^{\lambda}_{\text{dist}}(M, \mathcal{D}) := \mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\mathcal{D}}\Big[\phi^{\lambda}\big(\rho_{\mathrm{M}}(x_1,x_2), Y\big)\Big],$$

where $\phi^{\lambda}(\rho_{\mathrm{M}}, Y)$ is a generic distance-based loss function that computes the degree of violation between weighted distance $\rho_{\mathrm{M}}(x_1,x_2) := \|M(x_1 - x_2)\|^2$ and the label agreement $Y := \mathbf{1}[y_1 = y_2]$ and penalizes it by factor λ. For example, ϕ could penalize intra-class distances that are more than some upper limit U and inter-class distances that are less than some lower limit $L > U$:

$$\phi^{\lambda}_{L,U}(\rho_{\mathrm{M}}, Y) := \begin{cases} \min\{1, \lambda[\rho_{\mathrm{M}} - U]_+\} & \text{if } Y = 1 \\ \min\{1, \lambda[L - \rho_{\mathrm{M}}]_+\} & \text{otherwise} \end{cases}, \tag{1}$$

[1] Note that we are looking at the linear form of the metric M; usually the corresponding quadratic form $M^{\mathsf{T}}M$ is discussed in the literature, which is necessarily positive semi-definite.

where $[A]_+ := \max\{0, A\}$. MMC optimizes an efficiently computable variant of Eq. (1) by constraining the aggregate intra-class distances while maximizing the aggregate inter-class distances. ITML explicitly includes the upper and lower limits with an added regularization on the learned M to be close to a pre-specified metric of interest M_0.

While we will discuss loss-functions ϕ that handle distances between *pairs* of observations, it is easy to extend to relative distances among *triplets*:

$$\phi^\lambda_{\text{triple}}(\rho_{\text{M}}(x_1, x_2), \rho_{\text{M}}(x_1, x_3), (y_1, y_2, y_3)) := \begin{cases} \min\{1, \lambda[\rho_{\text{M}}(x_1, x_2) - \rho_{\text{M}}(x_1, x_3)]_+\} & \text{if } y_1 = y_2 \neq y_3 \\ 0 & \text{otherwise} \end{cases},$$

LMNN is a popular variant, in which instead of looking at all triplets, it focuses on triplets in local neighborhoods, improving the quality of local distance comparisons.

The **classifier-based framework** prefers metrics M that directly improve the prediction quality for a downstream task. Let $\mathcal{H}$ represent a real-valued hypothesis class associated with the prediction task of interest (each $h \in \mathcal{H} : X \to [0, 1]$), then the corresponding classifier-based error becomes:

$$\text{err}_{\text{hypoth}}(M, \mathcal{D}) := \inf_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\, \mathbf{1}\left[|h(Mx) - y| \geq 1/2\right]\Big].$$

Example classifier-based methods include [5], which minimizes ranking errors for information retrieval and [6], which incorporates network topology constraints for predicting network connectivity structure.

3 Metric Learning Sample Complexity: General Case

In any practical setting, we estimate the ideal weighting metric M^* by minimizing the empirical version of the error criterion from a finite size sample from $\mathcal{D}$. Let S_m denote a sample of size m, and $\text{err}(M, S_m)$ denote the corresponding empirical error. We can then define the empirical risk minimizing metric based on m samples as $M^*_m := \operatorname{argmin}_M \text{err}(M, S_m)$, and compare its generalization performance to that of the theoretically optimal M^*, that is,

$$\text{err}(M^*_m, \mathcal{D}) - \text{err}(M^*, \mathcal{D}). \qquad (2)$$

Distance-Based Error Analysis. Given an i.i.d. sequence of observations $z_1, z_2, \ldots$ from $\mathcal{D}$, $z_i = (x_i, y_i)$, we can pair the observations together to form a *paired* sample[2] $S^{\text{pair}}_m = \{(z_1, z_2), \ldots, (z_{2m-1}, z_{2m})\} = \{(z_{1,i}, z_{2,i})\}_{i=1}^m$ of size m, and define the sample-based distance error induced by a metric M as

$$\text{err}^\lambda_{\text{dist}}(M, S^{\text{pair}}_m) := \frac{1}{m}\sum_{i=1}^{m} \phi^\lambda\big(\,\rho_{\text{M}}(x_{1,i}, x_{2,i}), \mathbf{1}[y_{1,i} = y_{2,i}]\big).$$

Then for any B-bounded-support distribution $\mathcal{D}$ (that is, each $(x, y) \sim \mathcal{D}$, $\|x\| \leq B$), we have the following.[3,4]

Theorem 1 *Let ϕ^λ be a distance-based loss function that is λ-Lipschitz in the first argument. Then with probability at least $1 - \delta$ over an i.i.d. draw of $2m$ samples from an unknown B-bounded-support distribution $\mathcal{D}$ paired as S^{pair}_m, we have*

$$\sup_{M \in \mathcal{M}} \big[\, \text{err}^\lambda_{\text{dist}}(M, \mathcal{D}) - \text{err}^\lambda_{\text{dist}}(M, S^{\text{pair}}_m)\big] \leq O\left(\lambda B^2 \sqrt{D \ln(1/\delta)/m}\right).$$

[2] While we pair $2m$ samples into m independent pairs, it is common to consider all $O(m^2)$ possibly dependent pairs. By exploiting independence we provide a simpler analysis yielding $O(m^{-1/2})$ sample complexity rates, which is similar to the dependent case.

[3] We only present the results for paired comparisons; the results are easily extended to triplet comparisons.

[4] All the supporting proofs are provided in Appendix A.

This implies a bound on our key quantity of interest, Eq. (2). To achieve estimation error rate ϵ, $m = \Omega((\lambda B^2/\epsilon)^2 D \ln(1/\delta))$ samples are sufficient, showing that one never needs more than a number proportional to D examples to achieve the desired level of accuracy with high probability.

Since many applications involve high-dimensional data, we next study if such a strong dependency on D is necessary. It turns out that even for simple distance-based loss functions like $\phi^{\lambda}_{L,U}$ (c.f. Eq. 1), there are data distributions for which one cannot ensure good estimation error with fewer than linear in D samples.

Theorem 2 *Let $\mathcal{A}$ be any algorithm that, given an i.i.d. sample S_m (of size m) from a fixed unknown bounded support distribution $\mathcal{D}$, returns a weighting metric from $\mathcal{M}$ that minimizes the empirical error with respect to distance-based loss function $\phi^{\lambda}_{L,U}$. There exist $\lambda \geq 0$, $0 \leq U < L$ (indep. of D), s.t. for all $0 < \epsilon, \delta < \frac{1}{64}$, there exists a bounded support distribution $\mathcal{D}$, such that if $m \leq \frac{D+1}{512\epsilon^2}$,*

$$\mathbf{P}_{S_m}\left[\mathrm{err}^{\lambda}_{\text{dist}}(\mathcal{A}(S_m), \mathcal{D}) - \mathrm{err}^{\lambda}_{\text{dist}}(M^*, \mathcal{D}) > \epsilon \right] > \delta.$$

While this strong dependence on D may seem discouraging, note that here we made no assumptions about the underlying structure of the data distribution. One may be able to achieve a more relaxed dependence on D in settings in which individual features contain varying amounts of useful information. This is explored in Section 4.

Classifier-Based Error Analysis. In this setting, we consider an i.i.d. set of observations $z_1, z_2, \ldots$ from $\mathcal{D}$ to obtain the unpaired sample $S_m = \{z_i\}_{i=1}^m$ of size m. To analyze the generalization-ability of weighting metrics optimized w.r.t. underlying real-valued hypothesis class $\mathcal{H}$, we must measure the classification complexity of $\mathcal{H}$. The scale-sensitive version of VC-dimension, the *fat-shattering dimension*, of a hypothesis class (denoted $\mathsf{Fat}_\gamma(\mathcal{H})$) encodes the right notion of classification complexity and provides a way to relate generalization error to the empirical error at a *margin* γ [9].

In the context of metric learning with respect to a fixed hypothesis class, define the empirical error at a margin γ as $\mathrm{err}^{\gamma}_{\text{hypoth}}(M, S_m) := \inf_{h \in \mathcal{H}} \frac{1}{m} \sum_{(x_i, y_i) \in S_m} \mathbf{1}[\mathsf{Margin}(h(Mx_i), y_i) \leq \gamma]$, where $\mathsf{Margin}(\hat{y}, y) := (2y - 1)(\hat{y} - 1/2)$.

Theorem 3 *Let $\mathcal{H}$ be a λ-Lipschitz base hypothesis class. Pick any $0 < \gamma \leq 1/2$, and let $m \geq \mathsf{Fat}_{\gamma/16}(\mathcal{H}) \geq 1$. Then with probability at least $1 - \delta$ over an i.i.d. draw of m samples S_m from an unknown B-bounded-support distribution $\mathcal{D}$ ($\epsilon_0 := \min\{\gamma/2, 1/2\lambda B\}$)*

$$\sup_{M \in \mathcal{M}} \left[\mathrm{err}_{\text{hypoth}}(M, \mathcal{D}) - \mathrm{err}^{\gamma}_{\text{hypoth}}(M, S_m) \right] \leq O\left(\sqrt{\frac{1}{m} \ln \frac{1}{\delta} + \frac{D^2}{m} \ln \frac{D}{\epsilon_0} + \frac{\mathsf{Fat}_{\gamma/16}(\mathcal{H})}{m} \ln\left(\frac{m}{\gamma}\right)} \right).$$

As before, this implies a bound on Eq. (2). To achieve estimation error rate ϵ, $m = \Omega((D^2 \ln(\lambda D B/\gamma) + \mathsf{Fat}_{\gamma/16}(\mathcal{H}) \ln(1/\delta\gamma))/\epsilon^2)$ samples suffices. Note that the task of finding an optimal metric only additively increases sample complexity over that of finding the optimal hypothesis from the underlying hypothesis class. In contrast to the distance-based framework (Theorem 1), here we get a quadratic dependence on D. The following shows that a strong dependence on D is necessary in the absence of assumptions on the data distribution and base hypothesis class.

Theorem 4 *Pick any $0 < \gamma < 1/8$. Let $\mathcal{H}$ be a base hypothesis class of λ-Lipschitz functions that is closed under addition of constants (i.e., $h \in \mathcal{H} \implies h' \in \mathcal{H}$, where $h' : x \mapsto h(x) + c$, for all c) s.t. each $h \in \mathcal{H}$ maps into the interval $[1/2 - 4\gamma, 1/2 + 4\gamma]$ after applying an appropriate theshold. Then for any metric learning algorithm $\mathcal{A}$, and for any $B \geq 1$, there exists $\lambda \geq 0$, for all $0 < \epsilon, \delta < 1/64$, there exists a B-bounded-support distribution $\mathcal{D}$ s.t. if $m \ln^2 m < O\big(\frac{D^2 + d}{\epsilon^2 \ln(1/\gamma^2)}\big)$*

$$\mathbf{P}_{S_m \sim \mathcal{D}}[\mathrm{err}_{\text{hypoth}}(M^*, \mathcal{D}) > \mathrm{err}^{\gamma}_{\text{hypoth}}(\mathcal{A}(S_m), \mathcal{D}) + \epsilon] > \delta,$$

where $d := \mathsf{Fat}_{768\gamma}(\mathcal{H})$ is the fat-shattering dimension *of $\mathcal{H}$ at margin 768γ.*

4 Sample Complexity for Data with Un- and Weakly Informative Features

We introduce the concept of the *metric learning complexity* of a given dataset. Our key observation is that a metric that yields good generalization performance should emphasize relevant features while suppressing the contribution of spurious features. Thus, a good metric reflects the quality of individual feature measurements of data and their relative value for the learning task. We can leverage this and define the metric learning complexity of a given dataset as the *intrinsic complexity* d of the weighting metric that yields the best generalization performance for that dataset (if multiple metrics yield best performance, we select the one with minimum d). A natural way to characterize the intrinsic complexity of a weighting metric M is via the norm of the matrix M. Using metric learning complexity as our gauge for feature-set richness, we now refine our analysis in both canonical frameworks. We will first analyze sample complexity for norm-bounded metrics, then show how to *automatically adapt* to the intrinsic complexity of the unknown underlying data distribution.

4.1 Distance-Based Refinement

We start with the following refinement of the distance-based metric learning sample complexity for a class of Frobenius norm-bounded weighting metrics.

Lemma 5 *Let $\mathcal{M}$ be any class of weighting metrics on the feature space $X = \mathbb{R}^D$, and define $d := \sup_{M\in\mathcal{M}} \|M^{\mathsf{T}}M\|_F^2$. Let ϕ^λ be any distance-based loss function that is λ-Lipschitz in the first argument. Then with probability at least $1-\delta$ over an i.i.d. draw of $2m$ samples from an unknown B-bounded-support distribution $\mathcal{D}$ paired as S_m^{pair}, we have*

$$\sup_{M\in\mathcal{M}} \left[\mathrm{err}^{\lambda}_{\text{dist}}(M,\mathcal{D}) - \mathrm{err}^{\lambda}_{\text{dist}}(M,S^{\text{pair}}_m)\right] \le O\left(\lambda B^2 \sqrt{d\ln(1/\delta)/m}\right).$$

Observe that if our dataset has a low metric learning complexity $d \ll D$, then considering an appropriate class of norm-bounded weighting metrics $\mathcal{M}$ can help sharpen the sample complexity result, yielding a *dataset-dependent* bound. Of course, a priori we do not know which class of metrics is appropriate; We discuss how to *automatically adapt* to the right complexity class in Section 4.3.

4.2 Classifier-Based Refinement

Effective data-dependent analysis of classifier-based metric learning requires accounting for potentially complex interactions between an arbitrary base hypothesis class and the distortion induced by a weighting metric to the unknown underlying data distribution. To make the analysis tractable while still keeping our base hypothesis class $\mathcal{H}$ general, we assume that $\mathcal{H}$ is a class of two-layer feed-forward networks.[5] Recall that for any smooth target function f^*, a two-layer feed-forward neural network (with appropriate number of hidden units and connection weights) can approximate f^* arbitrarily well [10], so this class is flexible enough to include most reasonable target hypotheses.

More formally, define the base hypothesis class of two-layer feed-forward neural network with K hidden units as $\mathcal{H}^{\text{2-net}}_{\sigma^\gamma} := \{x \mapsto \sum_{i=1}^K w_i\ \sigma^\gamma(v_i\ \cdot\ x) \mid \|w\|_1 \le 1, \|v_i\|_1 \le 1\}$, where $\sigma^\gamma : \mathbb{R} \to [-1,1]$ is a smooth, strictly monotonic, γ-Lipschitz activation function with $\sigma^\gamma(0) = 0$. Then, for generalization error w.r.t. any classifier-based λ-Lipschitz loss function ϕ^λ,

$$\mathrm{err}^{\lambda}_{\text{hypoth}}(M,D) := \inf_{h\in\mathcal{H}^{\text{2-net}}_{\sigma^\gamma}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\phi^\lambda\big(h(Mx),y\big)\right],$$

we have the following.[6]

[5] We only present the results for two-layer networks in Lemma 6; the results are easily extended to multilayer feed-forward networks.

[6] Since we know the functional form of the base hypothesis class $\mathcal{H}$ (*i.e.,* a two layer feed-forward neural net), we can provide a more precise bound than leaving it as $\mathsf{Fat}(\mathcal{H})$.

Lemma 6 *Let $\mathcal{M}$ be any class of weighting metrics on the feature space $X = \mathbb{R}^D$, and define $d := \sup_{M \in \mathcal{M}} \|M^\mathsf{T} M\|_{_F}^2$. For any $\gamma > 0$, let $\mathcal{H}_{\sigma^\gamma}^{\text{2-net}}$ be a two layer feed-forward neural network base hypothesis class (as defined above) and ϕ^λ be a classifier-based loss function that λ-Lipschitz in its first argument. Then with probability at least $1 - \delta$ over an i.i.d. draw of m samples S_m from an unknown B-bounded support distribution $\mathcal{D}$, we have*

$$\sup_{M \in \mathcal{M}} \left[\operatorname{err}_{\text{hypoth}}^{\lambda}(M, \mathcal{D}) - \operatorname{err}_{\text{hypoth}}^{\lambda}(M, S_m) \right] \leq O\left(B\lambda\gamma \sqrt{d \ln(D/\delta)/m} \right).$$

4.3 Automatically Adapting to Intrinsic Complexity

While Lemmas 5 and 6 provide a sample complexity bound tuned to the metric learning complexity of a given dataset, these results are *not* directly useful since one cannot select the correct norm-bounded class $\mathcal{M}$ a priori, as the underlying distribution $\mathcal{D}$ is unknown. Fortunately, by considering an appropriate sequence of norm-bounded classes of weighting metrics, we can provide a uniform bound that *automatically adapts* to the intrinsic complexity of the unknown underlying data distribution $\mathcal{D}$.

Theorem 7 *Define $\mathcal{M}^d := \{M \mid \|M^\mathsf{T} M\|_{_F}^2 \leq d\}$, and consider the nested sequence of weighting metric classes $\mathcal{M}^1 \subset \mathcal{M}^2 \subset \cdots$. Let μ_d be any non-negative measure across the sequence $\mathcal{M}^d$ such that $\sum_d \mu_d = 1$ (for $d = 1, 2, \cdots$). Then for any $\lambda \geq 0$, with probability at least $1 - \delta$ over an i.i.d. draw of sample S_m from an unknown B-bounded-support distribution $\mathcal{D}$, for all $d = 1, 2, \cdots$, and all $M^d \in \mathcal{M}^d$,*

$$\left[\operatorname{err}^{\lambda}(M^d, \mathcal{D}) - \operatorname{err}^{\lambda}(M^d, S_m) \right] \leq O\left(C \cdot B\lambda \sqrt{d \ln(1/\delta\mu_d)/m} \right), \tag{3}$$

where $C := B$ for distance-based error, or $C := \gamma\sqrt{\ln D}$ for classifier-based error (for $\mathcal{H}_{\sigma^\gamma}^{\text{2-net}}$).

In particular, for a data distribution $\mathcal{D}$ that has metric learning complexity at most $d \in \mathbb{N}$, if there are $m \geq \Omega\left(d(CB\lambda)^2 \ln(1/\delta\mu_d)/\epsilon^2 \right)$ samples, then with probability at least $1 - \delta$

$$\left[\operatorname{err}^{\lambda}(M_m^{\text{reg}}, \mathcal{D}) - \operatorname{err}^{\lambda}(M^*, \mathcal{D}) \right] \; \leq \; O(\epsilon),$$

for $M_m^{\text{reg}} := \underset{M \in \mathcal{M}}{\operatorname{argmin}} \left[\operatorname{err}^{\lambda}(M, S_m) + \Lambda_{_M} d_{_M} \right]$, $\Lambda_{_M} := CB\lambda \sqrt{\ln(\delta\mu_{_{\lceil d_M^2 \rceil}})^{-1}/m}$, $d_{_M} := \|M^\mathsf{T} M\|_{_F}$.

The measure μ_d above encodes our prior belief on the complexity class $\mathcal{M}^d$ from which a target metric is selected by a metric learning algorithm given the training sample S_m. In absence of any prior beliefs, μ_d can be set to $1/D$ (for $d = 1, \ldots, D$) for scale constrained weighting metrics ($\sigma_{\max} = 1$). Thus, for an unknown underlying data distribution $\mathcal{D}$ with metric learning complexity d, with number of samples just proportional to d, we can find a good weighting metric.

This result also highlights that the generalization error of *any* weighting metric returned by an algorithm is proportional to the (smallest) norm-bounded class to which it belongs (cf. Eq. 3). If two metrics M_1 and M_2 have similar empirical errors on a given sample, but have different intrinsic complexities, then the expected risk of the two metrics can be considerably different. We expect the metric with lower intrinsic complexity to yield better generalization error. This partly explains the observed empirical success of norm-regularized optimization for metric learning [7, 8].

Using this as a guiding principle, we can design an improved optimization criteria for metric learning that jointly minimizes the sample error and a Frobenius norm regularization penalty. In particular,

$$\min_{M \in \mathcal{M}} \quad \operatorname{err}(M, S_m) \quad + \quad \Lambda \, \|M^\mathsf{T} M\|_{_F} \tag{4}$$

for any error criteria 'err' used in a downstream prediction task and a regularization parameter Λ. Similar optimizations have been studied before [7, 8], here we explore the practical efficacy of this augmented optimization on existing metric learning algorithms in high noise regimes where a dataset's intrinsic dimension is much smaller than its representation dimension.

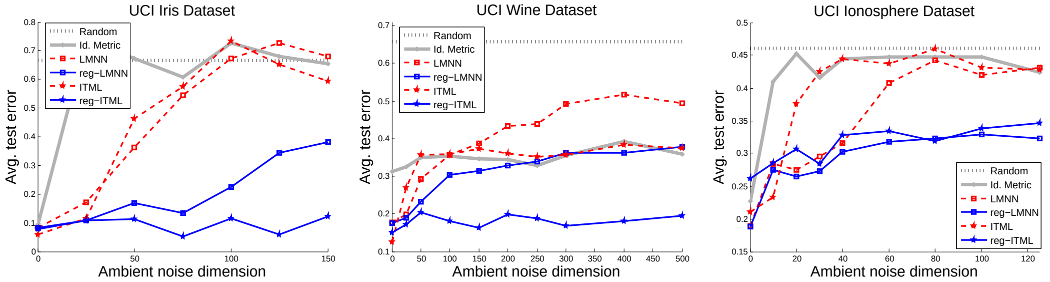

Figure 1: Nearest-neighbor classification performance of LMNN and ITML metric learning algorithms without regularization (dashed red lines) and with regularization (solid blue lines) on benchmark UCI datasets. The horizontal dotted line is the classification error of random label assignment drawn according to the class proportions, and solid gray line shows classification error of k-NN performance with respect to identity metric (no metric learning) for baseline reference.

5 Empirical Evaluation

Our analysis shows that the generalization error of metric learning can scale with the representation dimension, and regularization can help mitigate this by adapting to the intrinsic *metric learning complexity* of the given dataset. We want to explore to what degree these effects manifest in practice.

We select two popular metric learning algorithms, LMNN [1] and ITML [2], that are used to find metrics that improve nearest-neighbor classification quality. These algorithms have varying degrees of regularization built into their optimization criteria: LMNN implicitly regularizes the metric via its "large margin" criterion, while ITML allows for explicit regularization by letting the practitioners specify a "prior" weighting metric. We modified the LMNN optimization criteria as per Eq. (4) to also allow for an explicit norm-regularization controlled by the trade-off parameter Λ.

We can evaluate how the unregularized criteria (*i.e.,* unmodified LMNN, or ITML with the prior set to the identity matrix) compares to the regularized criteria (*i.e.,* modified LMNN with best Λ, or ITML with the prior set to a low-rank matrix).

Datasets. We use the UCI benchmark datasets for our experiments: IRIS (4 dim., 150 samples), WINE (13 dim., 178 samples) and IONOSPHERE (34 dim., 351 samples) datasets [11]. Each dataset has a fixed (unknown, but low) intrinsic dimension; we can vary the representation dimension by augmenting each dataset with synthetic correlated noise of varying dimensions, simulating regimes where datasets contain large numbers of uninformative features. Each UCI dataset is augmented with synthetic D-dimensional correlated noise as detailed in Appendix B.

Experimental setup. Each noise-augmented dataset was randomly split between 70% training, 10% validation, and 20% test samples. We used the default settings for each algorithm. For regularized LMNN, we picked the best performing trade-off parameter Λ from $\{0, 0.1, 0.2, ..., 1\}$ on the validation set. For regularized ITML, we seeded with the rank-one discriminating metric, *i.e.,* we set the prior as the matrix with all zeros, except the diagonal entry corresponding to the most discriminating coordinate set to one. All the reported results were averaged over 20 runs.

Results. Figure 1 shows the nearest-neighbor performance (with $k = 3$) of LMNN and ITML on noise-augmented UCI datasets. Notice that the unregularized versions of both algorithms (dashed red lines) scale poorly when noisy features are introduced. As the number of uninformative features grows, the performance of both algorithms quickly degrades to that of classification performance in the original unweighted space with no metric learning (solid gray line), showing poor adaptability to the signal in the data.

The regularized versions of both algorithms (solid blue lines) significantly improve the classification performance. Remarkably, regularized ITML shows almost no degradation in classification performance, even in very high noise regimes, demonstrating a strong robustness to noise. These results underscore the value of regularization in metric learning, showing that regularization encourages adaptability to the intrinsic complexity and improved robustness to noise.

6 Discussion and Related Work

Previous theoretical work on metric learning has focused almost exclusively on analyzing upper-bounds on the sample complexity in the distance-based framework, without exploring any intrinsic properties of the input data. Our work improves these results and additionally analyzes the classifier-based framework. It is, to best of our knowledge, the first to provide lower bounds showing that the dependence on D is necessary. Importantly, it is also the first to provide an analysis of sample rates based on a notion of intrinsic complexity of a dataset, which is particularly important in metric learning, where we expect the representation dimension to be much higher than intrinsic complexity.

[12] studied the norm-regularized convex losses for *stable* algorithms and showed an upper-bound sublinear in $\sqrt{D}$, which can be relaxed by applying techniques from [13]. We analyze the ERM criterion directly (thus no assumptions are made about the optimization algorithm), and provide a precise characterization of when the problem complexity is independent of D (Lm. 5). Our lower-bound (Thm. 2) shows that the dependence on D is necessary for ERM in the assumption-free case.

[14] and [15] analyzed the ERM criterion, and are most similar to our results providing an upper-bound for the distance-based framework. [14] shows a $O(m^{-1/2})$ rate for thresholds on bounded convex losses for distance-based metric learning without explicitly studying the dependence on D. Our upper-bound (Thm. 1) improves this result by considering arbitrary (possibly non-convex) distance-based Lipschitz losses and explicitly revealing the dependence on D. [15] provides an alternate ERM analysis of norm-regularized metrics and parallels our norm-bounded analysis in Lemma 5. While they focus on analyzing a specific optimization criterion (thresholds on the hinge loss with norm-regularization), our result holds for general Lipschitz losses. Our Theorem 7 extends it further by explicitly showing when we can expect good generalization performance from a given dataset.

[16] provides an interesting analysis for *robust* algorithms by relying upon the existence of a partition of the input space where each cell has similar training and test losses. Their sample complexity bound scales with the partition size, which in general can be exponential in D.

It is worth emphasizing that none of these closely related works discuss the importance of or leverage the intrinsic structure in data for the metric learning problem. Our results in Section 4 formalize an intuitive notion of dataset's intrinsic complexity for metric learning, and show sample complexity rates that are finely tuned to this *metric learning complexity*. Our lower bounds indicate that exploiting the structure is necessary to get rates that don't scale with representation dimension D.

The classifier-based framework we discuss has parallels with the kernel learning and similarity learning literature. The typical focus in kernel learning is to analyze the generalization ability of linear separators in Hilbert spaces [17, 18]. Similarity learning on the other hand is concerned about finding a similarity function (that does not necessarily has a positive semidefinite structure) that can best assist in linear classification [19, 20]. Our work provides a complementary analysis for learning explicit linear transformations of the given representation space for arbitrary hypotheses classes.

Our theoretical analysis partly justifies the empirical success of norm-based regularization as well. Our empirical results show that such regularization not only helps in designing new metric learning algorithms [7, 8], but can even benefit existing metric learning algorithms in high-noise regimes.

Acknowledgments

We would like to thank Aditya Menon for insightful discussions, and the anonymous reviewers for their detailed comments that helped improve the final version of this manuscript.

References

[1] K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research (JMLR)*, 10:207–244, 2009.

[2] J.V. Davis, B. Kulis, P. Jain, S. Sra, and I.S. Dhillon. Information-theoretic metric learning. *International Conference on Machine Learning (ICML)*, pages 209–216, 2007.

[3] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. *Neural Information Processing Systems (NIPS)*, 2004.

[4] E.P. Xing, A.Y. Ng, M.I. Jordan, and S.J. Russell. Distance metric learning with application to clustering with side-information. *Neural Information Processing Systems (NIPS)*, pages 505–512, 2002.

[5] B. McFee and G.R.G. Lanckriet. Metric learning to rank. *International Conference on Machine Learning (ICML)*, 2010.

[6] B. Shaw, B. Huang, and T. Jebara. Learning a distance metric from a network. *Neural Information Processing Systems (NIPS)*, 2011.

[7] D.K.H. Lim, B. McFee, and G.R.G. Lanckriet. Robust structural metric learning. *International Conference on Machine Learning (ICML)*, 2013.

[8] M.T. Law, N. Thome, and M. Cord. Fantope regularization in metric learning. *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] M. Anthony and P. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 1999.

[10] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 4:359–366, 1989.

[11] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[12] R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. *Neural Information Processing Systems (NIPS)*, pages 862–870, 2009.

[13] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research (JMLR)*, 2:499–526, 2002.

[14] W. Bian and D. Tao. Learning a distance metric by empirical loss minimization. *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1186–1191, 2011.

[15] Q. Cao, Z. Guo, and Y. Ying. Generalization bounds for metric and similarity learning. *CoRR*, abs/1207.5437, 2013.

[16] A. Bellet and A. Habrard. Robustness and generalization for metric learning. *CoRR*, abs/1209.1086, 2012.

[17] Y. Ying and C. Campbell. Generalization bounds for learning the kernel. *Conference on Computational Learning Theory (COLT)*, 2009.

[18] C. Cortes, M. Mohri, and A. Rostamizadeh. New generalization bounds for learning kernels. *International Conference on Machine Learning (ICML)*, 2010.

[19] M-F. Balcan, A. Blum, and N. Srebro. Improved guarantees for learning via similarity functions. *Conference on Computational Learning Theory (COLT)*, 2008.

[20] A. Bellet, A. Habrard, and M. Sebban. Similarity learning for provably accurate sparse linear classification. *International Conference on Machine Learning (ICML)*, 2012.

[21] Z. Guo and Y. Ying. Generalization classification via regularized similarity learning. *Neural Computation*, 26(3):497–552, 2014.

[22] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *CoRR*, abs/1306.6709, 2014.

[23] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research (JMLR)*, 3:463–482, 2002.

[24] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications*. 2010.

A Appendix: Various Proofs

A.1 Proof of Theorem 1

Let $\mathcal{P}$ be the probability measure induced by the random variable $(\mathbf{X}, Y)$, where $\mathbf{X} := (x, x')$, $Y := \mathbf{1}[y = y']$, st. $((x, y), (x', y')) \sim (\mathcal{D} \times \mathcal{D})$.

Define function class

$$\mathcal{F} := \left\{ f_M : \mathbf{X} \mapsto \|M(x - x')\|^2 \;\middle|\; \begin{array}{c} M \in \mathcal{M} \\ \mathbf{X} = (x, x') \in (X \times X) \end{array} \right\},$$

and consider any loss function $\phi^\lambda(\rho, Y)$ that is λ-Lipschitz in the first argument. Then, we are interested in bounding the quantity

$$\sup_{f_M \in \mathcal{F}} \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{P}}[\phi^\lambda(f_M(\mathbf{X}), Y)] - \frac{1}{m}\sum_{i=1}^m \phi^\lambda(f_M(\mathbf{X}_i), Y_i),$$

where $\mathbf{X}_i := (x_{1,i}, x_{2,i})$ and $Y_i := \mathbf{1}[y_{1,i} = y_{2,i}]$ and the sample based versions from the paired sample $S_m^{\text{pair}} = \{((x_{1,i}, y_{1,i}), (x_{2,i}, y_{2,i}))\}_{i=1}^m$.

Define $\bar{x}_i := x_{1,i} - x_{2,i}$ for each $\mathbf{X}_i = (x_{1,i}, x_{2,i})$. Then, the Rademacher complexity[7] of our function class $\mathcal{F}$ (with respect to the distribution $\mathcal{P}$) is bounded, since (let $\sigma_1, \ldots, \sigma_m$ denote independent uniform $\{\pm 1\}$-valued random variables)

$$\begin{aligned}
\mathcal{R}_m(\mathcal{F}, \mathcal{P}) &:= \mathbb{E}_{\substack{\mathbf{X}_i, \sigma_i \\ i \in [m]}}\left[\sup_{f_M \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f_M(\mathbf{X}_i)\right] \\
&= \frac{1}{m}\mathbb{E}_{\substack{\mathbf{X}_i, \sigma_i \\ i \in [m]}} \sup_{M \in \mathcal{M}} \left[\sum_{i=1}^m \sigma_i \bar{x}_i^\mathsf{T} M^\mathsf{T} M \bar{x}_i\right] \\
&= \frac{1}{m}\mathbb{E}_{\substack{\mathbf{X}_i, \sigma_i \\ i \in [m]}} \sup_{\substack{M \in \mathcal{M}, \text{ s.t.} \\ [a^{jk}]_{jk} := M^\mathsf{T} M}} \left[\sum_{j,k} a^{jk} \sum_{i=1}^m \sigma_i \bar{x}_i^j \bar{x}_i^k\right] \\
&\leq \frac{1}{m}\mathbb{E}_{\substack{\mathbf{X}_i, \sigma_i \\ i \in [m]}} \sup_{M \in \mathcal{M}} \left[\|M^\mathsf{T} M\|_{\text{F}} \left(\sum_{j,k}\Big(\sum_{i=1}^m \sigma_i \bar{x}_i^j \bar{x}_i^k\Big)^2\right)^{1/2}\right] \\
&\leq \frac{\sqrt{D}}{m}\mathbb{E}_{\mathbf{X}_i, i \in [m]}\left(\mathbb{E}_{\sigma_i, i \in [m]} \sum_{j,k}\Big(\sum_{i=1}^m \sigma_i \bar{x}_i^j \bar{x}_i^k\Big)^2\right)^{1/2} \\
&= \frac{\sqrt{D}}{m}\mathbb{E}_{\mathbf{X}_i, i \in [m]}\left(\sum_{j,k}\sum_{i=1}^m \Big(\bar{x}_i^j\Big)^2 \Big(\bar{x}_i^k\Big)^2\right)^{1/2} \\
&= \frac{\sqrt{D}}{m}\mathbb{E}_{\mathbf{X}_i, i \in [m]}\left(\sum_{i=1}^m \|\bar{x}_i\|^4\right)^{1/2} \\
&= \frac{\sqrt{D}}{m}\mathbb{E}_{\substack{(x_i, x_i') \sim (\mathcal{D}|_X \times \mathcal{D}|_X), \\ i \in [m]}}\left(\sum_{i=1}^m \|x_i - x_i'\|^4\right)^{1/2}
\end{aligned}$$

[7] See the definition of Rademacher complexity in the statement of Lemma 8.

$$\leq \sqrt{\frac{D}{m}} \left(\mathbb{E}_{(x,x') \sim (\mathcal{D}|_X \times \mathcal{D}|_X)} \|x - x'\|^4 \right)^{1/2}$$

$$\leq 4B^2 \sqrt{\frac{D}{m}},$$

where the second inequality is by noting that $\sup_{M \in \mathcal{M}} \|M^\mathsf{T} M\|_{_F} \leq \sqrt{D}$ for the class of weighting metrics $\mathcal{M} := \{M \mid M \in \mathbb{R}^{D \times D}, \sigma_{\max}(M) = 1\}$.

Recall that $\mathcal{D}$ has bounded support (with bound B). Thus, by noting that ϕ^λ is $8B^2$ bounded function that is λ-Lipschitz in the first argument, we can apply Lemma 8 and get the desired uniform deviation bound. ▮

Lemma 8 (Rademacher complexity of bounded Lipschitz loss functions [23]) *Let $\mathcal{D}$ be a fixed unknown distribution over $X \times \{-1, 1\}$, and let S_m be an i.i.d. sample of size m from $\mathcal{D}$. Given a hypothesis class $\mathcal{H} \subset \mathbb{R}^X$ and a loss function $\ell : \mathbb{R} \times \{-1, 1\} \to \mathbb{R}$, such that ℓ is c-bounded, and is λ-Lipschitz in the first argument, that is, $\sup_{(y',y) \in \mathbb{R} \times \{-1,1\}} |\ell(y', y)| \leq c$, and $|\ell(y', y) - \ell(y'', y)| \leq \lambda |y' - y''|$, we have the following:*

for any $0 < \delta < 1$, with probability at least $1 - \delta$, every $h \in \mathcal{H}$ satisfies

$$\text{err}(\ell \circ h, \mathcal{D}) \leq \text{err}(\ell \circ h, S_m) + 2\lambda \mathcal{R}_m(\mathcal{H}, \mathcal{D}) + c\sqrt{\frac{2\ln(1/\delta)}{m}},$$

where

- $\text{err}(\ell \circ h, \mathcal{D}) := \mathbb{E}_{x,y \sim \mathcal{D}}[\ell(h(x), y)]$,
- $\text{err}(\ell \circ h, S_m) := \frac{1}{m} \sum_{(x_i, y_i) \in S_m} \ell(h(x_i), y_i)$,
- *$\mathcal{R}_m(\mathcal{H}, \mathcal{D})$ is the Rademacher complexity of the function class $\mathcal{H}$ with respect to the distribution $\mathcal{D}$ given m i.i.d. samples, and is defined as:*

$$\mathcal{R}_m(\mathcal{H}, \mathcal{D}) := \mathbb{E}_{\substack{x_i \sim \mathcal{D}|_X, \\ \sigma_i \sim \text{unif}\{\pm 1\}, \\ i \in [m]}} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right],$$

where σ_i are independent uniform $\{\pm 1\}$-valued random variables.

A.2 Proof of Theorem 2

We shall exhibit a finite class of bounded support distributions $\mathfrak{D}$, such that if $\mathcal{D}$ is chosen uniformly at random from $\mathfrak{D}$, the expectation (over the random choice of $\mathcal{D}$) of the probability of failure (that is, generalization error of the metric returned by $\mathcal{A}$ that uses fewer than $O(D/\epsilon^2)$ samples compared to that of the optimal metric exceeds the specified tolerance level ϵ) is at least δ. This implies that for some distribution in $\mathfrak{D}$ the probability of failure is at least δ as well, yielding the desired result.

Let $\Delta_D := \{x_0, \ldots, x_D\}$ be a set of $D + 1$ points that form the vertices of a regular simplex (that is circumscribed in a $(D - 1)$-sphere of radius $D/4$) from the underlying space $X = \mathbb{R}^D$ as per Definition 1 (see below). For a fixed parameter $0 < \alpha < 1$ (exact value to be determined later), define $\mathfrak{D}'$ as the class of all distributions $\mathcal{D}$ on $X \times \{0, 1\}$ such that:

- $\mathcal{D}$ assigns zero probability to all sets not intersecting $\Delta_D \times \{0, 1\}$.
- for each $i = 0, \ldots, D$, either

- $\mathbf{P}[(x_i, 1)] = (1 + \sqrt{\alpha})/2$ and $\mathbf{P}[(x_i, 0)] = (1 - \sqrt{\alpha})/2$, or
- $\mathbf{P}[(x_i, 1)] = (1 - \sqrt{\alpha})/2$ and $\mathbf{P}[(x_i, 0)] = (1 + \sqrt{\alpha})/2$.

Observe that the class $\mathfrak{D}'$ contains 2^{D+1} distributions. We shall discard two distributions from this class: the distribution that assigns $P[(x_i, 1)] = (1+\sqrt{\alpha})/2$ for all i, and the distribution that assigns $P[(x_i, 1)] = (1-\sqrt{\alpha})/2$ for all i. This reduced collection of $2^{D+1}-2$ distributions will be denoted by $\mathfrak{D}$.

For concreteness, we shall use a specific instantiation of $\phi^{\lambda}_{L,U}$ in $\text{err}^{\lambda}_{\text{dist}}$ with $U = 0$, $L = 1$ and $\lambda = 1$, in our proof below.

Proof overview. We first show, by the construction of the distributions under consideration in $\mathfrak{D}$, the sample error and the generalization error minimizing metrics over any $\mathcal{D} \in \mathfrak{D}$ belong to a restricted class of weighting matrices (Eq. 5). We then make a second simplification by noting that finding these (sample- and generalization-) error minimizing metrics (in the restricted class) is equivalent to solving a binary classification problem (Eq. 6). This reduction to binary classification enables us to use VC-style lower bounding techniques to give a lower bound on the sample complexity. We now fill in the details.

Consider a subset of weighting metrics $\mathcal{M}_{\text{0-1}}$ that map points in Δ_D to exactly one of two possible points that are (squared) distance at least 1 apart, that is,

$$\begin{aligned}\mathcal{M}_{\text{0-1}} := \{M \mid M \in \mathcal{M}, \exists z_0, z_1 \in \mathbb{R}^D, \forall x \in \Delta_D,\\ Mx \in \{z_0, z_1\} \text{ and } \|z_0 - z_1\|^2 \geq 1\}.\end{aligned}$$

Now pick any $\mathcal{D} \in \mathfrak{D}$, and let $S_m := (x_1, y_1), \ldots, (x_m, y_m)$ be an i.i.d. sample of size m from $\mathcal{D}$, and denote the corresponding paired sample as S^{pair}_m. Observe that both the sample-based and the distribution-based error minimizing weighting metric from $\mathcal{M}$ on $\mathcal{D}$ also belongs to $\mathcal{M}_{\text{0-1}}$. That is, (c.f. Lemma 10 and our choice of U, L and λ)

$$\begin{aligned}\text{argmin}_{M\in\mathcal{M}}\ \text{err}^{\lambda}_{\text{dist}}(M, \mathcal{D}) \in \mathcal{M}_{\text{0-1}}\\ \text{argmin}_{M\in\mathcal{M}}\ \text{err}^{\lambda}_{\text{dist}}(M, S^{\text{pair}}_m) \in \mathcal{M}_{\text{0-1}}.\end{aligned} \tag{5}$$

A reduction to binary classification on product space. For each $M \in \mathcal{M}_{\text{0-1}}$, we associate a classifier $f_M : (\Delta_D \times \Delta_D) \to \{0, 1\}$ defined as $(x_i, x_j) \mapsto \mathbf{1}[Mx_i = Mx_j]$. Now, consider the probability measure $\mathcal{P}$ induced by the random variable $(\mathbf{X}, Y)$, where $\mathbf{X} := (x, x')$, $Y := \mathbf{1}[y = y']$, s.t. $((x, y), (x', y')) \sim \big(\mathcal{D}|_{(\Delta_D\times\{0,1\})} \times \mathcal{D}|_{(\Delta_D\times\{0,1\})}\big)$. It is easy to check that for all $M \in \mathcal{M}_{\text{0-1}}$

$$\begin{aligned}\text{err}^{\lambda}_{\text{dist}}(M, \mathcal{D}) &= \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{P}}\big[\,\mathbf{1}[f_M(\mathbf{X}) \neq Y]\big]\\ \text{err}^{\lambda}_{\text{dist}}(M, S^{\text{pair}}_m) &= \frac{1}{|S^{\text{pair}}_m|} \sum_{((x,y),(x',y'))\in S^{\text{pair}}_m} \mathbf{1}\big[f_M((x, x')) \neq \mathbf{1}[y = y']\big].\end{aligned} \tag{6}$$

Define

$$\begin{aligned}\eta(\mathbf{X}) \;&:=\; \mathbf{P}_{Y\sim\mathcal{P}|_{Y|\mathbf{X}}}[Y = 1|\,\mathbf{X}]\\ &=\; \mathbf{P}_{(y,y')\sim(\mathcal{D}\times\mathcal{D})|_{(y,y')|(x,x')}}[y = y'|x, x']\\ &=\; \begin{cases} \frac{1}{2} + \frac{\alpha}{2} & \text{if } \mathbf{P}(y|x) = \mathbf{P}(y'|x') \\ \frac{1}{2} - \frac{\alpha}{2} & \text{if } \mathbf{P}(y|x) \neq \mathbf{P}(y'|x') \end{cases}.\end{aligned} \tag{7}$$

Observe that $\eta(\mathbf{X})$ is the Bayes error rate at $\mathbf{X}$ for distribution $\mathcal{P}$. Since, by construction of $\mathcal{M}_{\text{0-1}}$, the class $\{f_M\}_{M\in\mathcal{M}_{\text{0-1}}}$ contains a classifier that achieves the Bayes error rate, the optimal classifier

$f^* := \operatorname{argmin}_{f_M} \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{P}} \mathbf{1}[f_M(\mathbf{X}) \neq Y]$ necessarily has $f^*(\mathbf{X}) = \mathbf{1}[\eta(\mathbf{X}) > \frac{1}{2}]$ (for all $\mathbf{X}$). Then, for any f_M,

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{P}}\big[\mathbf{1}[f_M(\mathbf{X}) \neq Y]\big] - \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{P}}\big[\mathbf{1}[f^*(\mathbf{X}) \neq Y]\big] \\
&\qquad = \mathbb{E}_{\mathbf{X}\sim\mathcal{P}|_{\mathbf{X}}}\big[\eta(\mathbf{X})\big(\mathbf{1}[f^*(\mathbf{X})=1] - \mathbf{1}[f_M(\mathbf{X})=1]\big) \\
&\qquad\qquad\qquad + (1-\eta(\mathbf{X}))\big(\mathbf{1}[f^*(\mathbf{X})=0] - \mathbf{1}[f_M(\mathbf{X})=0]\big)\big] \\
&\qquad = \mathbb{E}_{\mathbf{X}\sim\mathcal{P}|_{\mathbf{X}}}\big[(2\eta(\mathbf{X})-1)\big(\mathbf{1}[f^*(\mathbf{X})=1] - \mathbf{1}[f_M(\mathbf{X})=1]\big)\big] \\
&\qquad = \mathbb{E}_{\mathbf{X}\sim\mathcal{P}|_{\mathbf{X}}}\big[2|\eta(\mathbf{X})-1/2| \cdot \mathbf{1}[f_M(\mathbf{X}) \neq f^*(\mathbf{X})]\big] \\
&\qquad = \frac{2\alpha}{(D+1)^2} \sum_{0\le i<j\le D} \big[\mathbf{1}[f_M((x_i,x_j)) \neq f^*((x_i,x_j))]\big], \qquad (8)
\end{aligned}$$

where (i) the second to last equality is by noting that $f^*(\mathbf{X}) \neq 1 \iff \eta(\mathbf{X}) \le 1/2$, and (ii) the last equality is by noting Eq. (7), $f_M((x_i,x_i)) = f^*((x_i,x_i)) = 1$ for all i and $f((x_i,x_j)) = f((x_j,x_i))$ for all f. For notational simplicity, we shall define $\mathbf{X}_{i,j} := (x_i,x_j)$.

Now, for a given sample S_m, let $N(S_m) := (N_i)_i$ (for all $0 \le i \le D$), where N_i is the number of occurrences of the point x_i in S_m. Then for any f_M,

$$\begin{aligned}
&\mathbb{E}_{S_m}\left[\frac{1}{(D+1)^2}\sum_{i<j}\mathbf{1}[f_M(\mathbf{X}_{i,j}) \neq f^*(\mathbf{X}_{i,j})]\right] \\
&\quad = \frac{1}{(D+1)^2}\sum_{i<j}\mathbf{P}_{S_m}[f_M(\mathbf{X}_{i,j}) \neq f^*(\mathbf{X}_{i,j})] \\
&\quad = \frac{1}{(D+1)^2}\sum_{i<j}\sum_{N\in\mathbb{N}^{D+1}}\mathbf{P}_{S_m}[f_M(\mathbf{X}_{i,j}) \neq f^*(\mathbf{X}_{i,j}) | N(S_m) = N] \cdot \mathbf{P}[N(S_m) = N] \\
&\quad = \frac{1}{(D+1)^2}\sum_{N\in\mathbb{N}^{D+1}}\mathbf{P}[N(S_m) = N] \cdot \sum_{i<j}\mathbf{P}_{S_m}[f_M(\mathbf{X}_{i,j}) \neq f^*(\mathbf{X}_{i,j}) | N_i, N_j].
\end{aligned}$$

For any algorithm $\mathcal{A}$ that takes the sample S_m as an input and returns a metric M that minimizes the empirical error, we have

$$\begin{aligned}
&\mathbb{E}_{S_m}\left[\frac{1}{(D+1)^2}\sum_{i<j}\mathbf{1}[f_{\mathcal{A}(S_m)}(\mathbf{X}_{i,j}) \neq f^*(\mathbf{X}_{i,j})]\right] \\
&\quad = \frac{1}{(D+1)^2}\sum_{N\in\mathbb{N}^{D+1}}\mathbf{P}[N(S_m) = N] \cdot \sum_{i<j}\mathbf{P}_{S_m}[f_{\mathcal{A}(S_m)}(\mathbf{X}_{i,j}) \neq f^*(\mathbf{X}_{i,j}) | N_i, N_j] \\
&\quad \ge \frac{1}{(D+1)^2}\sum_{N\in\mathbb{N}^{D+1}}\mathbf{P}[N(S_m) = N] \cdot \sum_{i<j}\frac{1}{4}\left(1 - \sqrt{1 - \exp\left(\frac{-(\max\{N_i,N_j\}+1)\alpha^2}{1-\alpha^2}\right)}\right) \\
&\quad \ge \frac{1}{4}\frac{D}{D+1}\left(1 - \sqrt{1 - \exp\left(\frac{-(m/(D+1)+1)\alpha^2}{1-\alpha^2}\right)}\right) \\
&\quad \ge \frac{1}{8}\left(1 - \sqrt{1 - \exp\left(\frac{-(m/(D+1)+1)\alpha^2}{1-\alpha^2}\right)}\right),
\end{aligned}$$

where (i) the first inequality is by applying Lemma 11, (ii) the second inequality is by assuming WLOG $N_i \ge N_j$, and noting that the expression above is convex in N_i so one can apply Jensen's inequality and by observing that $\mathbb{E}[N_i] = m/(D+1)$ and that there are total $D(D+1)$ summands for $i < j$, and (iii) the last inequality is by noting that $D \geq 1$. Now, let B denote the r.h.s. quantity above. Then by recalling that for any $[0,1]$-valued random variable Z, $\mathbf{P}(Z > \gamma) > \mathbb{E}Z - \gamma$ (for all $0 < \gamma < 1$), we have

$$\mathbf{P}_{S_m}\Big[\frac{1}{(D+1)^2}\sum_{i<j}\mathbf{1}\left[f_{\mathcal{A}(S_m)}((x_i,x_j)) \neq f^*((x_i,x_j))\right] > \gamma B\Big] > (1-\gamma)B.$$

Or equivalently, by combining Eqs. (5), (6) and (8), we have

$$\mathbb{E}_{\mathcal{D}\sim\text{unif}(\mathfrak{D})}\mathbf{P}_{S_m\sim\mathcal{D}}\Big[\text{err}^{\lambda}_{\text{dist}}(\mathcal{A}(S_m),\mathcal{D}) - \text{err}^{\lambda}_{\text{dist}}(M^*_{\mathcal{D}},\mathcal{D}) > 2\alpha\gamma B\Big] > (1-\gamma)B,$$

where $M^*_{\mathcal{D}} := \text{argmin}_{M\in\mathcal{M}}\,\text{err}^{\lambda}_{\text{dist}}(M,\mathcal{D})$ and $\mathcal{A}(S_m)$ is any metric returned by empirical error minimizing algorithm. Now, if (cond. 1) $B \geq \delta/1-\gamma$ and (cond. 2) $\epsilon \leq 2\gamma\alpha B$ hold, it follows that for some $\mathcal{D} \in \mathfrak{D}$

$$\mathbf{P}_{S_m\sim\mathcal{D}}\Big[\text{err}^{\lambda}_{\text{dist}}(\mathcal{A}(S_m),\mathcal{D}) - \text{err}^{\lambda}_{\text{dist}}(M^*_{\mathcal{D}},\mathcal{D}) > \epsilon\Big] > \delta. \tag{9}$$

To satisfy cond. 1 & 2, we shall select $\gamma = 1 - 16\delta$. Then cond. 1 follows if

$$m \leq (D+1)\left(\frac{1-\alpha^2}{\alpha^2}\ln(4/3) - 1\right).$$

Choosing parameter $\alpha = 8\epsilon/\gamma$ (and by noting $B \geq 1/16$ by cond. 1 for choice of γ and m), cond. 2 is satisfied as well. Hence,

$$m \leq (D+1)\left(\frac{(1-16\delta)^2 - (8\epsilon)^2}{64\epsilon^2}\ln(4/3) - 1\right)$$

implies Eq. (9). Moreover, if $0 < \epsilon, \delta < 1/64$ then $m \leq \frac{(D+1)}{512\epsilon^2}$ would suffice. ▌

Definition 1 *Define $D+1$ vectors $\Delta_D := \{v_0, \ldots, v_D\}$, with each $v_i \in \mathbb{R}^D$ as*

$$v_{0,j} := -1/2 \qquad\qquad \textit{for } 1 \leq j \leq D$$

$$v_{i,j} := \begin{cases} \frac{(D-1)\sqrt{D+1}+1}{2D} & \textit{if } i = j \\ \frac{1-\sqrt{D+1})}{2D} & \textit{otherwise} \end{cases} \qquad \textit{for } 1 \leq i,j \leq D$$

Fact 9 (properties of vertices of a regular D-simplex) *Let $\Delta_D = \{v_0, \ldots, v_D\}$ be a set of $D+1$ vectors in $\mathbb{R}^D$ as per Definition 1. Then, Δ_D defines vertices of a regular D-simplex circumscribed in a $(D-1)$-sphere of radius $\sqrt{D}/2$, with*

(i) $\|v_i\|^2 = D/4$ *(for all i), and*

(ii) $\|v_i - v_j\|^2 = (D+1)/2$ *(for $i \neq j$).*

Moreover, for any non-empty bi-partition of Δ_D into $\Delta^{(1)}_D$ and $\Delta^{(2)}_D$ with $|\Delta^{(1)}_D| = k$ and $|\Delta^{(2)}_D| = D+1-k$, define $a^{(1)}$ and $a^{(2)}$ the means (centroids) of the points in $\Delta^{(1)}_D$ and $\Delta^{(2)}_D$ respectively. Then, we also have

(i) $(a^{(1)} - a^{(2)}) \cdot (a^{(i)} - v_j) = 0$ *(for $i \in \{1,2\}$, and $v_j \in \Delta^{(i)}_D$).*

(ii) $\|a^{(1)} - a^{(2)}\|^2 = \frac{(D+1)^2}{4k(D+1-k)} \geq 1$, *for* $1 \leq k \leq D$.

Lemma 10 *Let* Δ_D *be a set of* $D+1$ *points* $\{X_0, \dots, X_D\}$ *in* $\mathbb{R}^D$ *as per Definition 1, and let* $\mathcal{D}$ *be an arbitrary distribution over* $\Delta_D \times \{0,1\}$. *Define the following quantities:*

- $P_i := \mathbf{1}[\mathbf{P}_{\mathcal{D}}[(X_i, 1)] > 1/2]$ *(for* $0 \leq i \leq D$*).*
- $\Pi := \{\pi : \Delta_D \to \mathbb{R}^D\}$ *as the collection of all functions that map points in* Δ_D *to arbitrary points in* $\mathbb{R}^D$.
- $f((x,y),(x',y');\pi) := \begin{cases} \min\{1, \|\pi(x) - \pi(x')\|^2\} & \textit{if } y = y' \\ \min\{1, [1 - \|\pi(x) - \pi(x')\|^2]_+\} & \textit{if } y \neq y' \end{cases}$ *as the pairwise loss function with respect to the mapping function* $\pi \in \Pi$.
- $\mathcal{E}(\pi) := \mathbb{E}_{(x,y),(x',y') \sim \mathcal{D} \times \mathcal{D}}[f((x,y),(x',y');\pi)]$ *as the average loss induced by a mapping* π, *and* $\mathcal{E}^* := \inf_{\pi \in \Pi} \mathcal{E}(\pi)$ *as the minimum possible achievable error.*

Then, for any $\bar{\pi} \in \Pi$ *such that*

(i) $\bar{\pi}(X_i) = \bar{\pi}(X_j)$, *if* $P_i = P_j$

(ii) $\|\bar{\pi}(X_i) - \bar{\pi}(X_j)\|^2 \geq 1$, *if* $P_i \neq P_j$,

we have that $\mathcal{E}(\bar{\pi}) = \mathcal{E}^*$. *Moreover, define* $\bar{A}$ *as*

- $\bar{A} := \frac{A_1 - A_0}{\|A_1 - A_0\|}$, *where* $A_0 := \text{mean}(X_i)$ *such that* $P_i = 0$, *and* $A_1 := \text{mean}(X_i)$ *such that* $P_i = 1$ *(if exists at least one* $P_i = 0$ *and at least one* $P_i = 1$*).*
- $\bar{A} := 0$, *i.e. the zero vector in* $\mathbb{R}^D$ *(otherwise).*

And let M *be a* $D \times D$ *matrix defined as*

$$M := \bar{A}\bar{A}^{\mathsf{T}}.$$

Then, we have the following:

(i) either M *is identically the zero matrix (in the case when* $\bar{A} = 0$*), or the maximum singular value of* M, $\sigma_{\max}(M) = 1$.

(ii) the linear map $\pi_M : x \mapsto Mx$ *satisfies conditions (i) and (ii) above, and thus* $\mathcal{E}(\pi_M) = \mathcal{E}^*$.

Proof. The proof follows from the geometric properties of the vertices of a regular simplex Δ_D and Fact 9. ▮

Lemma 11 *Given two random variables* α_1 *and* α_2, *each uniformly distributed on* $\{\alpha_-, \alpha_+\}$ *independently, where* $\alpha_- = 1/2 - \epsilon/2$ *and* $\alpha_+ = 1/2 + \epsilon/2$ *with* $0 < \epsilon < 1$. *Suppose that* $\xi_1^1, \dots, \xi_{m_1}^1$ *and* $\xi_1^2, \dots, \xi_{m_2}^2$ *are two i.i.d. sequences of* $\{0,1\}$*-valued random variables with* $\mathbf{P}(\xi_i^1 = 1) = \alpha_1$ *and* $\mathbf{P}(\xi_i^2 = 1) = \alpha_2$ *for all* i. *Then, for any likelihood maximizing function* f *from* $\{0,1\}^{\mathbb{N}}$ *to* $\{\alpha_-, \alpha_+\}$ *that estimates the bias* α_1 *and* α_2 *from the samples,*

$$\mathbf{P}\Big[\big(f(\xi_1^1, \dots, \xi_{m_1}^1) \neq \alpha_1 \text{ and } f(\xi_1^2, \dots, \xi_{m_2}^2) = \alpha_2\big),$$

$$\text{or}\big(f(\xi_1^1, \dots, \xi_{m_1}^1) = \alpha_1 \text{ and } f(\xi_1^2, \dots, \xi_{m_2}^2) \neq \alpha_2\big)\Big] > \frac{1}{4}\left(1 - \sqrt{1 - \exp\Big(\frac{-2\lceil \max\{m_1, m_2\}/2 \rceil \epsilon^2}{1 - \epsilon^2}\Big)}\right).$$

Proof. Note that

$$\begin{aligned}
&\mathbf{P}\Big[\big(f(\xi_1^1,\ldots,\xi_{m_1}^1)\neq\alpha_1 \text{ and } f(\xi_1^2,\ldots,\xi_{m_2}^2)=\alpha_2\big) \text{ or } \big(f(\xi_1^1,\ldots,\xi_{m_1}^1)=\alpha_1 \text{ and } f(\xi_1^2,\ldots,\xi_{m_2}^2)\neq\alpha_2\big)\Big]\\
&\quad= \mathbf{P}[f(\xi_1^1,\ldots,\xi_{m_1}^1)\neq\alpha_1]\cdot\mathbf{P}[f(\xi_1^2,\ldots,\xi_{m_2}^2)=\alpha_2]+\mathbf{P}[f(\xi_1^1,\ldots,\xi_{m_1}^1)=\alpha_1]\cdot\mathbf{P}[f(\xi_1^2,\ldots,\xi_{m_2}^2)\neq\alpha_2]\\
&\quad\geq \frac{1}{2}\mathbf{P}\big[f(\xi_1^1,\ldots,\xi_{m_1}^1)\neq\alpha_1\big]+\frac{1}{2}\mathbf{P}\big[f(\xi_1^2,\ldots,\xi_{m_2}^2)\neq\alpha_2\big]\\
&\quad> \frac{1}{4}\left(1-\sqrt{1-\exp\left(\frac{-2\lceil\max\{m_1,m_2\}/2\rceil\epsilon^2}{1-\epsilon^2}\right)}\,\right),
\end{aligned}$$

where the first inequality is by noting that a likelihood maximizing f will select the correct bias better than random (which has probability $1/2$), and the second inequality is by applying Lemma 12. ▮

Lemma 12 (Lemma 5.1 of [9]) *Suppose that α is a random variable uniformly distributed on $\{\alpha_-,\alpha_+\}$, where $\alpha_-=1/2-\epsilon/2$ and $\alpha_+=1/2+\epsilon/2$, with $0<\epsilon<1$. Suppose that $\xi_1,\ldots,\xi_m$ are i.i.d. $\{0,1\}$-valued random variables with $\mathbf{P}(\xi_i=1)=\alpha$ for all i. Let f be any function from $\{0,1\}^m$ to $\{\alpha_-,\alpha_+\}$. Then*

$$\mathbf{P}\big[f(\xi_1,\ldots,\xi_m)\neq\alpha\big]>\frac{1}{4}\left(1-\sqrt{1-\exp\Big(\frac{-2\lceil m/2\rceil\epsilon^2}{1-\epsilon^2}\Big)}\right).$$

A.3 Proof of Theorem 3

For any $M\in\mathcal{M}$ define real-valued hypothesis class on domain X as $\mathcal{H}_M:=\{x\mapsto h(Mx):h\in\mathcal{H}\}$ and define

$$\mathcal{F}:=\{x\mapsto h(Mx):M\in\mathcal{M},h\in\mathcal{H}\}=\bigcup_M\mathcal{H}_M.$$

Observe that a uniform convergence of errors induced by the functions in $\mathcal{F}$ implies convergence of the class of weighted matrices as well.

Now for any domain X, real-valued hypothesis class $\mathcal{G}\subset[0,1]^X$, margin $\gamma>0$, and a sample $S\subset X$, define

$$\mathrm{cov}_\gamma(\mathcal{G},S):=\left\{C\subset\mathcal{G}\;\middle|\;\begin{array}{c}\forall g\in\mathcal{G},\exists g'\in C,\\ \max_{s\in S}|g(s)-g'(s)|\leq\gamma\end{array}\right\}$$

as the set of γ-covers of S by $\mathcal{G}$. Let γ-covering number of $\mathcal{G}$ for any integer $m>0$ be defined as

$$\mathcal{N}_\infty(\gamma,\mathcal{G},m):=\max_{S\subset X:|S|=m}\;\min_{C\in\mathrm{cov}_\gamma(\mathcal{G},S)}|C|,$$

with the minimizing cover C called as the minimizing (γ,m)-cover of $\mathcal{G}$

Now, for the given γ, we will first estimate the γ-covering number of $\mathcal{F}$, that is, $\mathcal{N}_\infty(\gamma,\mathcal{F},m)$.

For any $M\in\mathcal{M}$, let H_M be the minimizing $(\gamma/2,m)$-cover of $\mathcal{H}_M$. Note that $|H_M|=\mathcal{N}_\infty(\gamma/2,\mathcal{H}_M,m)\leq\mathcal{N}_\infty(\gamma/2,\mathcal{H},m)$ (because $MX\subset X$).

Now let $\mathcal{M}_\epsilon$ be an ϵ-spectral cover of $\mathcal{M}$ (that is, for every $M\in\mathcal{M}$, exists $M'\in\mathcal{M}_\epsilon$ such that $\sigma_{\max}(M-M')\leq\epsilon$), and define

$$\bar{F}_\epsilon:=\{x\mapsto h(Mx):M\in\mathcal{M}_\epsilon,h\in\mathcal{H}_M\}.$$

Note that $|\bar{F}_\epsilon| \leq |\mathcal{M}_\epsilon||H_I| \leq \mathcal{N}_\infty(\gamma/2, \mathcal{H}, m)(1 + 2D/\epsilon)^{D^2}$ (c.f. Lemma 13). Observe that $\bar{F}_\epsilon$ is a $(\gamma/2 + B\lambda\epsilon)$-cover of $\mathcal{F}$, since (i) for any $f \in F$ (formed by combining, say, $M_0 \in \mathcal{M}$ and $h_0 \in \mathcal{H}$), exists $\bar{f} \in \bar{F}_\epsilon$, namely the $\bar{f}$ formed by $\bar{M}_0$ such that $\sigma_{\max}(M_0 - \bar{M}_0) \leq \epsilon$, and (ii) $\bar{h}_0 \in H_{\bar{M}_0}$ such that $|h_0(\bar{M}_0 x) - \bar{h}_0(\bar{M}_0 x)| \leq \gamma/2$ (for all $x \in X$). So, (for any $x \in X$)

$$\begin{aligned}
|f(x) - \bar{f}(x)| &= |h_0(M_0 x) - \bar{h}_0(\bar{M}_0 x)| \\
&\leq |h_0(M_0 x) - h_0(\bar{M}_0 x)| \\
&\quad + |h_0(\bar{M}_0 x) - \bar{h}_0(\bar{M}_0 x)| \\
&\leq \lambda \|M_0 x - \bar{M}_0 x\| + \gamma/2 \\
&\leq \lambda \sigma_{\max}(M_0 - \bar{M}_0)\|x\| + \gamma/2 \\
&\leq \lambda \epsilon B + \gamma/2.
\end{aligned}$$

So, if we pick $\epsilon = \min\{\frac{1}{2\lambda B}, \frac{\gamma}{2}\}$, it follows that

$$\mathcal{N}_\infty(\gamma, \mathcal{F}, m) \leq |\bar{F}_\epsilon| \leq \mathcal{N}_\infty(\gamma/2, \mathcal{H}, m)(1 + 2D/\epsilon)^{D^2}.$$

By noting Lemmas 14 and 15, it follows that

$$\begin{aligned}
&\mathbf{P}_{S_m \sim \mathcal{D}}\Big[\exists f \in \mathcal{F} : \mathrm{err}(f) \geq \mathrm{err}_\gamma(f, S_m) + \alpha\Big] \\
&\qquad\qquad \leq 4\Big(1 + \frac{2D}{\epsilon}\Big)^{D^2}\Big(\frac{128m}{\gamma^2}\Big)^{\mathsf{Fat}_{\gamma/16}(\mathcal{H}) \ln\left(\frac{32em}{\mathsf{Fat}_{\gamma/16}(\mathcal{H})\gamma}\right)} e^{-\alpha^2 m/8}.
\end{aligned}$$

The lemma follows by bounding this failure probability with at most δ. ∎

Lemma 13 (ϵ-spectral coverings of $D \times D$ matrices) *Let $\mathcal{M} := \{M \mid M \in \mathbb{R}^{D \times D}, \sigma_{\max}(M) = 1\}$ be the set of matrices with unit spectral norm. Define $\mathcal{M}_\epsilon$ as the ϵ-cover of $\mathcal{M}$, that is, for every $M \in \mathcal{M}$, there exists $M' \in \mathcal{M}_\epsilon$ such that $\sigma_{\max}(M - M') \leq \epsilon$. Then for all $\epsilon > 0$, there exists $\mathcal{M}_\epsilon$ such that $|\mathcal{M}_\epsilon| \leq \big(1 + \frac{2D}{\epsilon}\big)^{D^2}$.*

Proof. Fix any $\epsilon > 0$ and let $\mathcal{N}_{\epsilon/D}$ be a minimal size (ϵ/D)-cover of Euclidean unit ball $\mathbf{B}_D$ in $\mathbb{R}^D$. That is, for any $v \in \mathbf{B}_D$, there exists $v' \in \mathcal{N}_{\epsilon/D}$ such that $\|v - v'\| \leq \epsilon/D$. Using standard volume arguments (see e.g. proof of Lemma 5.2 of [24]), we know that $|\mathcal{N}_{\epsilon/D}| \leq \big(1 + \frac{2D}{\epsilon}\big)^D$. Define

$$\mathcal{M}_\epsilon := \Big\{M' \mid M' = [v'_1 \ \cdots \ v'_D] \in \mathbb{R}^{D \times D}, v'_i \in \mathcal{N}_{\epsilon/D}\Big\}.$$

Then $\mathcal{M}_\epsilon$ constitutes as an ϵ-cover of $\mathcal{M}$, since for any $M = [v_1 \cdots v_D] \in \mathcal{M}$ there exists $M' = [v'_1 \cdots v'_D] \in \mathcal{M}_\epsilon$, in particular M' such that $\|v_i - v'_i\| \leq \epsilon/D$ (for all i). Then

$$\sigma_{\max}(M - M') \leq \|M - M'\|_F = \sum_i \|v_i - v'_i\| \leq \epsilon.$$

Without loss of generality we can assume that each $M' \in \mathcal{M}_\epsilon$, $\sigma_{\max}(M') = 1$. Moreover, by construction, $|\mathcal{M}_\epsilon| \leq \big(1 + \frac{2D}{\epsilon}\big)^{D^2}$. ∎

Lemma 14 (extension of Theorem 12.8 of [9]) *Let $\mathcal{H}$ be a set of real functions from a domain X to the interval $[0, 1]$. Let $\gamma > 0$. Then for all $m \geq 1$,*

$$\mathcal{N}_\infty(\gamma, \mathcal{H}, m) < c_0(4m/\gamma^2)^{\mathsf{Fat}_{\gamma/4}(\mathcal{H}) \ln \frac{4em}{\mathsf{Fat}_{\gamma/4}(\mathcal{H})\gamma}}.$$

for some universal constant c_0.

Proof. Theorem 12.8 of [9] asserts this for $m \geq \mathsf{Fat}_{\gamma/4}(\mathcal{H}) \geq 1$ with $c_0 = 2$. Now, if $1 \leq m < \mathsf{Fat}_{\gamma/4}(\mathcal{H})$, for some universal constant c', we have $\mathcal{N}_\infty(\gamma, \mathcal{H}, m) \leq (c'/\gamma)^m \leq (c'/\gamma)^{\mathsf{Fat}_{\gamma/4}(\mathcal{H})}$. ▮

Lemma 15 (Theorem 10.1 of [9]) *Suppose that $\mathcal{H}$ is a set of real-valued functions defined on domain X. Let $\mathcal{D}$ be any probability distribution on $Z = X \times \{0,1\}$, $0 \leq \epsilon \leq 1$, real $\gamma > 0$ and integer $m \geq 1$. Then,*

$$\mathbf{P}_{S_m \sim \mathcal{D}}\Big[\exists h \in \mathcal{H} : \mathrm{err}(h) \geq \mathrm{err}_\gamma(h, S_m) + \epsilon\Big] \leq 2\mathcal{N}_\infty\Big(\frac{\gamma}{2}, \mathcal{H}, 2m\Big)e^{-\epsilon^2 m/8},$$

where S_m is an i.i.d. sample of size m from $\mathcal{D}$.

A.4 Proof of Theorem 4

For any fixed $0 < \gamma < 1/8$ and the given bounded class of distributions with bound $B \geq 1$, consider a $(1/B)$-bi-Lipschitz base hypothesis class $\mathcal{H}$ that maps hypothesis from the domain X to $[1/2 - 4\gamma, 1/2 + 4\gamma]$, and define

$$\mathcal{F} := \{x \mapsto h(Mx) : M \in \mathcal{M}, h \in \mathcal{H}\}.$$

Note that finding M that minimizes $\mathrm{err}_{\mathsf{hypoth}}$ is equivalent to finding f that minimizes error on $\mathcal{F}$. Using Lemma 19, we have for any $0 < \gamma < 1/2$, the sample complexity of $\mathcal{F}$ is (for all $0 < \epsilon, \delta < 1/64$)

$$m \geq \frac{\mathsf{Fat}_{2\gamma}(\pi_{4\gamma}(\mathcal{F}))}{320\epsilon^2}, \tag{10}$$

where $\pi_{4\gamma}(\mathcal{F})$ is the (4γ)-*squashed* function class of $\mathcal{F}$ (see Definition 2 below). We lower bound $\mathsf{Fat}_{2\gamma}(\pi_{4\gamma}(\mathcal{F}))$ in terms of fat-shattering dimension of $\mathcal{H}$ to yield the lemma.

To this end we shall first define the (γ, m)-covering and packing number of a generic real-valued hypothesis class $\mathcal{G}$. For any domain X, real-valued hypothesis class $\mathcal{G} \subset [0,1]^X$, margin $\gamma > 0$, and a sample $S \subset X$, define

$$\mathrm{cov}_\gamma(\mathcal{G}, S) := \left\{ C \subset \mathcal{G} \middle| \begin{array}{c} \forall g \in \mathcal{G}, \exists g' \in C, \\ \max_{s \in S} |g(s) - g'(s)| \leq \gamma \end{array} \right\},$$

$$\mathrm{pak}_\gamma(\mathcal{G}, S) := \left\{ P \subset \mathcal{G} \middle| \begin{array}{c} \forall g \neq g' \in P, \\ \max_{s \in S} |g(s) - g'(s)| \geq \gamma \end{array} \right\}$$

as the set of γ-covers (resp. γ-packings) of S by $\mathcal{G}$. Let γ-covering number (resp. γ-packing number) of $\mathcal{G}$ for any integer $m > 0$ be defined as

$$\mathcal{N}_\infty(\gamma, \mathcal{G}, m) := \max_{S \subset X : |S| = m} \ \min_{C \in \mathrm{cov}_\gamma(\mathcal{G}, S)} |C|,$$

$$\mathcal{P}_\infty(\gamma, \mathcal{G}, m) := \max_{S \subset X : |S| = m} \ \max_{P \in \mathrm{pak}_\gamma(\mathcal{G}, S)} |P|$$

with the minimizing cover C (resp. maximizing packing P) called as the minimizing (γ, m)-cover (resp. maximizing (γ, m)-packing) of $\mathcal{G}$.

With these definitions, we have the following (for some universal constant c_0).

$$\begin{aligned} c_0\Big(\frac{m}{16\gamma^2}\Big)^{\mathsf{Fat}_{2\gamma}(\pi_{4\gamma}(\mathcal{F}))\ln(em/2\gamma)} &\geq \mathcal{N}_\infty(8\gamma, \pi_{4\gamma}(\mathcal{F}), m) && \text{[Lemma 14]} \\ &\geq \mathcal{P}_\infty(16\gamma, \pi_{4\gamma}(\mathcal{F}), m) && \text{[Lemma 17]} \end{aligned}$$

$$
\begin{aligned}
&\geq \Big(\frac{1}{32\gamma}\Big)^{D^2} \mathcal{P}_\infty(48\gamma, \pi_{4\gamma}(\mathcal{H}), m) && \text{[see (*) below]} \\
&= \Big(\frac{1}{32\gamma}\Big)^{D^2} \mathcal{P}_\infty(48\gamma, \mathcal{H}, m) && \text{[by the choice of } \mathcal{H}\text{]} \\
&\geq \Big(\frac{1}{32\gamma}\Big)^{D^2} \mathcal{N}_\infty(48\gamma, \mathcal{H}, m) && \text{[Lemma 17]} \\
&\geq \Big(\frac{1}{32\gamma}\Big)^{D^2} e^{\mathsf{Fat}_{768\gamma}(\mathcal{H})/8}. && \text{[Lemma 18]} \quad (11)
\end{aligned}
$$

(*) We show that $\mathcal{P}_\infty(16\gamma, \pi_{4\gamma}(\mathcal{F}), m) \geq (1/32\gamma)^{D^2} \mathcal{P}_\infty(48\gamma, \pi_{4\gamma}(\mathcal{H}), m)$, by exhibiting a set $\mathcal{S} \subset \pi_{4\gamma}(\mathcal{F})$ of size $(1/32\gamma)^{D^2} \mathcal{P}_\infty(48\gamma, \pi_{4\gamma}(\mathcal{H}), m)$ that is a (16γ)-packing of $\pi_{4\gamma}(\mathcal{F})$.

Let $\pi_{4\gamma}(\mathcal{H}_{48\gamma}) \subset \pi_{4\gamma}(\mathcal{H})$ be a maximal (32γ)-packing of $\pi_{4\gamma}(\mathcal{H})$ (that is, a maximal set such that for all distinct $(\pi_{4\gamma} \circ h), (\pi_{4\gamma} \circ h') \in \pi_{4\gamma}(\mathcal{H}_{48\gamma})$, exists $x \in X$ such that $|\pi_{4\gamma}(h(x)) - \pi_{4\gamma}(h'(x))| \geq 48\gamma$). Fix ϵ (exact value determined later), and define

$$
\mathcal{S}_\epsilon := \left\{ x \mapsto (\pi_{4\gamma} \circ h)(Mx) \;\middle|\; \begin{array}{c} (\pi_{4\gamma} \circ h) \in \pi_{4\gamma}(\mathcal{H}_{48\gamma}), \\ M \in \mathcal{M}_\epsilon \end{array} \right\},
$$

where $\mathcal{M}_\epsilon$ is a ϵ-spectral net of $\mathcal{M}$, that is, for all $M \in \mathcal{M}$, exists $M' \in \mathcal{M}_\epsilon$ such that $\sigma_{\max}(M - M') \leq \epsilon$, and for all distinct $M', M'' \in \mathcal{M}_\epsilon$, $\sigma_{\max}(M' - M'') \geq \epsilon/2$.

Then for any two distinct $f, f' \in \mathcal{S}_\epsilon$, such that $f(x) = (\pi_{4\gamma} \circ h)(Mx)$ and $f'(x) = (\pi_{4\gamma} \circ h')(M'x)$, we have

- (case 1) h and h' are distinct. In this case, there exists $x \in X$, s.t.
$$
\begin{aligned}
|f(x) - f'(x)| = &|\pi_{4\gamma}(h(Mx)) - \pi_{4\gamma}(h'(M'x))| \\
&\geq |\pi_{4\gamma}(h(Mx)) - \pi_{4\gamma}(h'(Mx))| \\
&\quad - |\pi_{4\gamma}(h'(Mx)) - \pi_{4\gamma}(h'(M'x))| \\
&\geq 48\gamma - (1/B)\sigma_{\max}(M - M')\|x\| \\
&\geq 48\gamma - (1/B)\epsilon B = 48\gamma - \epsilon.
\end{aligned}
$$

- (case 2) h, h' same but M and M' distinct. In this case, there exists x (with $\|x\| = 1$) s.t.
$$
\begin{aligned}
|f(x) - f'(x)| &= |\pi_{4\gamma}(h(Mx)) - \pi_{4\gamma}(h(M'x))| \\
&= |h(Mx) - h(M'x)| \\
&\geq B\|(M - M')x\| \\
&\geq B \cdot \min_{M \neq M' \in \mathcal{M}_\epsilon} \sigma_{\max}(M - M') \\
&\geq B(\epsilon/2).
\end{aligned}
$$

Thus, by setting $\epsilon = 32\gamma$, distinct classifiers $f, f' \in \mathcal{S}_{32\gamma}$ are at least 16γ apart (since $B \geq 1$). Hence $\mathcal{S}_{32\gamma}$ forms a (16γ)-packing of $\pi_{4\gamma}(\mathcal{F})$. Therefore, the packing number

$$
\mathcal{P}_\infty(16\gamma, \pi_{4\gamma}(\mathcal{F}), m) \geq |\mathcal{S}_{32\gamma}| = |\mathcal{M}_{32\gamma}||\mathcal{H}_{48\gamma}| \geq (1/32\gamma)^{D^2} \mathcal{P}_\infty(48\gamma, \pi_{4\gamma}(\mathcal{H}), m).
$$

Thus, from Eq. (11), it follows that

$$
\mathsf{Fat}_{2\gamma}(\pi_{4\gamma}(\mathcal{F})) \geq \Omega\Big(\frac{D^2 \ln(1/\gamma) + \mathsf{Fat}_{768\gamma}(\mathcal{H})}{\ln(m/\gamma^2)\ln(m/\gamma)}\Big).
$$

Combining this with Eq. (10), the lemma follows. ▮

Lemma 16 **(ϵ-spectral packings of $D \times D$ matrices)** *Let $\mathcal{M} := \{M \mid M \in \mathbb{R}^{D\times D}, \sigma_{\max}(M) = 1\}$ be the set of matrices with unit spectral norm. Define $\mathcal{M}_\epsilon \subset \mathcal{M}$ as the ϵ-packing of $\mathcal{M}$, that is, for every distinct $M, M' \in \mathcal{M}_\epsilon$, $\sigma_{\max}(M - M') \geq \epsilon$. Then for all $\epsilon > 0$, there exists $\mathcal{M}_\epsilon$ such that $|\mathcal{M}_\epsilon| \geq \left(\frac{1}{2\epsilon}\right)^{D^2}$.*

Proof. Fix any $\epsilon > 0$ and let $\mathcal{P}_\epsilon$ be a maximal size ϵ-packing of Euclidean unit ball $\mathbf{B}_D$ in $\mathbb{R}^D$. That is, for all distinct $v, v' \in \mathbf{B}_D$, $\|v - v'\| \geq \epsilon$. Using standard volume arguments (see e.g. proof of Lemma 5.2 of [24]), we know that $|\mathcal{P}_\epsilon| \geq \left(\frac{1}{2\epsilon}\right)^D$. Define

$$\mathcal{M}_\epsilon := \Big\{ M' \bigm| M' = [v'_1 \ \cdots \ v'_D] \in \mathbb{R}^{D\times D}, v'_i \in \mathcal{P}_\epsilon \Big\}.$$

Then $\mathcal{M}_\epsilon$ constitutes as an ϵ-packing of $\mathcal{M}$, since for any distinct $M, M' \in \mathcal{M}_\epsilon$ such that $M = [v_1 \cdots v_D]$ and $M' = [v'_1 \cdots v'_D]$, we have

$$\sigma_{\max}(M - M') \geq \max_i \|v_i - v'_i\| \geq \epsilon.$$

Without loss of generality we can assume that each $M \in \mathcal{M}_\epsilon$, $\sigma_{\max}(M) = 1$. Moreover, by construction, $|\mathcal{M}_\epsilon| \geq \left(\frac{1}{2\epsilon}\right)^{D^2}$. ▮

Lemma 17 **(follows from Theorem 12.1 of [9])** *For any real valued hypothesis class $\mathcal{H}$ into $[0, 1]$, all $m \geq 1$, and $0 < \gamma < 1/2$,*

$$\mathcal{P}_\infty(2\gamma, \mathcal{H}, m) \leq \mathcal{N}_\infty(\gamma, \mathcal{H}, m) \leq \mathcal{P}_\infty(\gamma, \mathcal{H}, m).$$

Lemma 18 **(Theorem 12.10 of [9])** *Let $\mathcal{H}$ be a set of real functions from a domain X to the interval $[0, 1]$. Let $\gamma > 0$. Then for $m \geq \mathsf{Fat}_{16\gamma}(\mathcal{H})$,*

$$\mathcal{N}_\infty(\gamma, \mathcal{H}, m) \geq e^{\mathsf{Fat}_{16\gamma}(\mathcal{H})/8}.$$

Lemma 19 **(Theorem 13.5 of [9])** *Suppose that $\mathcal{H}$ is a class of real valued functions that is closed under addition of constants, that is,*

$$h \in \mathcal{H} \implies h' \in \mathcal{H}, \text{ where } h' : x \mapsto h(x) + c \quad \text{for all } c.$$

such that each $h \in \mathcal{H}$ maps into the interval $[0, 1]$ after applying an appropriate threshold. Pick any $0 < \gamma < 1/2$. Then for any classification learning algorithm $\mathcal{A}$ for $\mathcal{H}$, and for all $0 < \epsilon, \delta < 1/64$, there exists a distribution $\mathcal{D}$ such that if $m \leq \frac{d}{320\epsilon^2}$, then

$$\mathbf{P}_{S_m \sim \mathcal{D}}[\text{err}(h^*, \mathcal{D}) > \text{err}_\gamma(\mathcal{A}(S_m), \mathcal{D}) + \epsilon] > \delta$$

*where $d := \mathsf{Fat}_{2\gamma}(\pi_{4\gamma}(\mathcal{H})) \geq 1$ is the fat-shattering dimension of $\pi_{4\gamma}(\mathcal{H})$—the (4γ)-*squashed *function class of $\mathcal{H}$, see Definition 2 below—at margin 2γ.*

Definition 2 **(squashing function)** *For any $0 < \gamma < 1/2$, define the* squashing function *$\pi_\gamma : \mathbb{R} \to [1/2 - \gamma, 1/2 + \gamma]$ as*

$$\pi_\gamma(\alpha) = \begin{cases} 1/2 + \gamma & \textit{if } \alpha \geq 1/2 + \gamma \\ 1/2 - \gamma & \textit{if } \alpha \leq 1/2 - \gamma \\ \alpha & \textit{otherwise} \end{cases}.$$

Moreover, for a collection F of functions into $\mathbb{R}$, define $\pi_\gamma(F) := \{\pi_\gamma \circ f \mid f \in F\}$.

A.5 Proof of Lemma 5

Let $\mathcal{P}$ be the probability measure induced by the random variable $(\mathbf{X}, Y)$, where $\mathbf{X} := (x, x')$, $Y := \mathbf{1}[y = y']$, st. $((x,y),(x',y')) \sim (\mathcal{D} \times \mathcal{D})$.

Define function class

$$\mathcal{F} := \left\{ f_M : \mathbf{X} \mapsto \|M(x - x')\|^2 \,\middle|\, \begin{array}{c} M \in \mathcal{M} \\ \mathbf{X} = (x, x') \in (X \times X) \end{array} \right\},$$

Following the steps of proof of Theorem 1, we can conclude that the Rademacher complexity of $\mathcal{F}$ is bounded. In particular,

$$\mathcal{R}_m(\mathcal{F}) \leq 4B^2 \sqrt{\frac{\sup_{M \in \mathcal{M}} \|M^\mathsf{T} M\|_{_F}^2}{m}}.$$

The result follows by noting that ϕ is λ-Lipschitz in the first argument and by applying Lemma 8. ▌

A.6 Proof of Lemma 6

Consider the function class

$$\mathcal{F} := \Big\{ f_{v,M} : x \mapsto v \cdot Mx \;\big|\; \|v\|_1 \leq 1, M \in \mathcal{M} \Big\},$$

and define the composition class

$$\mathcal{F}_\sigma := \left\{ x \mapsto \sum_{i=1}^{K} w_i \sigma^\gamma(f_i(x)) \,\middle|\, \begin{array}{c} \|w_i\|_1 \leq 1, \\ f_1, \ldots, f_K \in \mathcal{F} \end{array} \right\}.$$

Then, first note that the Gaussian complexity of $\mathcal{F}$ (with respect to the distribution $\mathcal{D}$) is bounded, since (let $g_1, \ldots, g_m$ denote independent standard Gaussian random variables)

$$\begin{aligned}
\mathcal{G}_m(\mathcal{F}, \mathcal{D}) &:= \mathbb{E}_{\substack{x_i \sim \mathcal{D}|_X \\ g_i, i \in [m]}} \left[\sup_{f_{v,M} \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} g_i f_{v,M}(x_i) \right] \\
&= \frac{1}{m} \mathbb{E}_{\substack{x_i \sim \mathcal{D}|_X \\ g_i, i \in [m]}} \left[\sup_{\substack{M \in \mathcal{M} \\ \|v\|_1 \leq 1}} v \cdot \sum_{i=1}^{m} g_i (Mx_i) \right] \\
&= \frac{1}{m} \mathbb{E}_{\substack{x_i \sim \mathcal{D}|_X \\ g_i, i \in [m]}} \left[\max_j \sup_{M \in \mathcal{M}} \sum_{i=1}^{m} g_i (Mx_i)_j \right] \\
&\leq \frac{1}{m} \mathbb{E}_{\substack{x_i \sim \mathcal{D}|_X \\ g_i, i \in [m]}} \max_{j \in [D]} \left[\sum_{i=1}^{m} g_i \sup_{M \in \mathcal{M}} \big|(Mx_i)_j\big| \right] \\
&\leq \frac{c \ln^{\frac{1}{2}}(D)}{m} \mathbb{E}_{x_i \sim \mathcal{D}_X} \max_{j, j' \in [D]} \left(\mathbb{E}_{g_i} \left[\sum_{i=1}^{m} g_i \Big(\sup_{M \in \mathcal{M}} \big|(Mx_i)_j\big| - \sup_{M' \in \mathcal{M}} \big|(M'x_i)_{j'}\big| \Big) \right]^2 \right)^{\frac{1}{2}} \\
&= \frac{c \ln^{\frac{1}{2}}(D)}{m} \mathbb{E}_{x_i \sim \mathcal{D}_X} \max_{j, j' \in [D]} \left(\sum_{i=1}^{m} \Big[\sup_{M \in \mathcal{M}} \big|(Mx_i)_j\big| - \sup_{M' \in \mathcal{M}} \big|(M'x_i)_{j'}\big| \Big]^2 \right)^{\frac{1}{2}} \\
&\leq c' B \sqrt{\frac{d \ln D}{m}},
\end{aligned}$$

where (i) second to last inequality is by applying Lemma 20, (ii) c, c' are absolute constants, (iii) $d := \sup_{M \in \mathcal{M}} \|M^{\mathsf{T}} M\|_F^2$. Note that bounding the Gaussian complexity also bounds the Rademacher complexity by Lemma 21.

Finally by noting that $\mathcal{F}_\sigma$ is a γ-Lipschitz composition class of $\mathcal{F}$ and ϕ^λ is a classification based loss function that is λ-Lipschitz in the first argument, we can apply Lemma 8 yielding the desired result. ▮

Lemma 20 (Lemma 20 of [23]) *Let $Z_1, \ldots, Z_D$ be random variables such that each $Z_j = \sum_{i=1}^m a_{ij} g_i$, where each g_i is independent $N(0,1)$ random variables. Then there is an absolute constant c such that*

$$\mathbb{E}_{g_i} \max_j Z_j \le c \ln^{\frac{1}{2}}(D) \max_{j,j'} \sqrt{\mathbb{E}_{g_i}(Z_j - Z_{j'})^2}.$$

Lemma 21 (Lemma 4 of [23]) *There are absolute constants c and C such that for every class $\mathcal{F}$ and every integer m*

$$c\mathcal{R}_m(\mathcal{F}, \mathcal{D}) \;\le\; \mathcal{G}_m(\mathcal{F}, \mathcal{D}) \;\le\; C \ln(m) \mathcal{R}_m(\mathcal{F}, \mathcal{D}),$$

where $\mathcal{R}$ and $\mathcal{G}$ are Rademacher and Gaussian complexities of a function class $\mathcal{F}$ with respect to the distribution $\mathcal{D}$ respectively.

A.7 Proof of Theorem 7

The conclusion of Eq. (3) is immediate by dividing the given failure probability δ across the sequence $\mathcal{M}^1, \mathcal{M}^2, \cdots$ such that $\delta\mu_d$ failure probability is associated with class $\mathcal{M}^d$, then apply Lemma 5 (for distance based metric learning) or Lemma 6 (for classifier based metric learning) to each class $\mathcal{M}^d$ individually, and finally combining the individual deviations together with a union bound.

For the second part, for any $M \in \mathcal{M}$ define d_M and Λ_M as per the lemma statement. Then with probability at least $1 - \delta$

$$\begin{aligned}
\mathrm{err}^\lambda(M_m^{\mathrm{reg}}, \mathcal{D}) - \mathrm{err}^\lambda(M^*, \mathcal{D}) &\le \mathrm{err}^\lambda(M_m^{\mathrm{reg}}, S_m) + d_{M_m^{\mathrm{reg}}} \Lambda_{M_m^{\mathrm{reg}}} - \mathrm{err}^\lambda(M^*, \mathcal{D}) \\
&\le \mathrm{err}^\lambda(M^*, S_m) + d_{M^*} \Lambda_{M^*} - \mathrm{err}^\lambda(M^*, \mathcal{D}) \\
&\le O(d_{M^*} \Lambda_{M^*}) \;=\; O(\epsilon),
\end{aligned}$$

where (i) the first inequality is by applying Eq. (3) on weighting metric M_m^{reg} (with failure probability set to $\delta/2$), (ii) the second inequality is by noting that M_m^{reg} is the (regularized) sample error minimizer as per the lemma statement, (iii) the third inequality is by applying Eq. (3) on weighting metric M^* (with failure probability set to $\delta/2$), and (iv) the last equality by noting the definitions of Λ_{M^*} and our choice of m. ▮

B Appendix: Creating Correlated-Synthetic-Noise Augmented Dataset

We first sample a covariance matrix Σ_D from unit-scale Wishart distribution (that is, let A be a $D \times D$ Gaussian random matrix with entry $A_{ij} \sim N(0,1)$ drawn i.i.d., and set $\Sigma_D := A^{\mathsf{T}} A$). Then each sample x_i from the dataset is appended independently by drawing noise vector $x_\sigma \sim N(0, \Sigma_D)$.

We varied the ambient noise dimension D between 0 and 500 dimensions and added it to the UCI datasets, creating the noise-augmented datasets.