Sample Complexity of Rank Regression Using Pairwise Comparisons

Berkan Kadıoğlu[a,1,*], Peng Tian[a,1,], Jennifer Dy[a], Deniz Erdoğmuş[a], Stratis Ioannidis[a]

[a]*Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, 02115, USA.*

Abstract

We consider a rank regression setting, in which a dataset of N samples with features in $\mathbb{R}^d$ is ranked by an oracle via M pairwise comparisons. Specifically, there exists a latent total ordering of the samples; when presented with a pair of samples, a noisy oracle identifies the one ranked higher with respect to the underlying total ordering. A learner observes a dataset of such comparisons and wishes to regress sample ranks from their features. We show that to learn the model parameters with $\epsilon > 0$ accuracy, it suffices to conduct $M \in \Omega(dN \log^3 N/\epsilon^2)$ comparisons uniformly at random when N is $\Omega(d/\epsilon^2)$.

Keywords: sample complexity, rank regression, pairwise comparisons, features.

1. Introduction

Rank regression has a broad range of applications in fields as diverse as social science [1, 2, 3], economics [4, 5], and medicine [6, 7, 8], to name a few. For example, disease severity can be regressed from patient records by presenting pairs to a medical expert and asking her to rank them [6]. A dataset of such pairwise comparisons is more informative than a dataset with class labels containing diagnostic outcomes. This is because comparisons reveal *intra-class, relative* severity within, e.g., the healthy or diseased class, that cannot be inferred from class labels alone. As an additional practical benefit, comparison labels also often exhibit lower variability across experts: experts are more likely to agree when comparing pairs rather than making absolute diagnoses: this has been observed in a variety of domains, including medicine [9, 10, 11], movie recommendations [12], travel recommendations [2], music recommendations [3], and web page recommendations [1]. These advantages make learning from comparisons quite advantageous in practice; in an extreme example illustrating this, Yıldız et al. [7] used comparisons among just 80 images to train a neural network of 5.9 million parameters that attained a 0.92 AUC on a much larger test set.

*Corresponding author

Email addresses: kadioglu.b@ece.neu.edu (Berkan Kadıoğlu), pengtian@ece.neu.edu (Peng Tian), jdy@ece.neu.edu (Jennifer Dy), erdogmus@ece.neu.edu (Deniz Erdoğmuş), ioannidis@ece.neu.edu (Stratis Ioannidis)

[1]Berkan Kadıoğlu and Peng Tian are both first authors of this paper.

This empirical success motivates us to study the sample complexity of algorithms that learn from comparisons. However, doing so poses a significant challenge. In contrast to the standard probably approximately correct (PAC) learning setting, where samples are assumed to be i.i.d, learning from comparisons necessarily leads to a violation of independence. Even in a simple generative model where (a) samples are drawn independently and (b) pairs presented to the oracle are selected uniformly at random, any two pairs sharing a sample are correlated. This dependence complicates the application of concentration inequalities such as, e.g., Chernoff bounds in this setting.

The main contributions of our work are as follows. We propose an estimator for the parameters of a generalized linear parametric model, which encompasses classical preference models such as Bradley-Terry [13] and Thurstone [14]. We overcome the aforementioned violation of independence and prove a sample complexity guarantee on model parameters. In particular, assuming Gaussian distributed features, we characterize the convergence of the estimator to a rescaled version of the model parameters w.r.t. the ambient dimension d, the number of samples N, and the number of comparisons M presented to the oracle. We show that to attain an accuracy $\epsilon > 0$ in model parameters, it suffices to conduct $\Omega(dN \log^3 N/\epsilon^2)$ comparisons when the number of samples is $\Omega(d/\epsilon^2)$. Finally, we confirm this dependence with experiments on synthetic data.

2. Related Work

In *rank aggregation* [15, 16], subsets of samples are ranked by a noisy oracle, and a learner attempts to reconstruct a total ordering from these noisy rankings without access to sample features. Works on noisy sorting assume that the observed pairwise comparisons deviate from an existing underlying ordering via i.i.d. Bernoulli noise. Braverman and Mossel [17] propose a tractable active learning algorithm that requires $\Omega(N \log(N))$ comparisons to recover the underlying ordering with high probability. Jamieson and Nowak [18] actively rank samples with $\Omega(d \log^2 N)$ pairwise comparisons when samples are embedded into an unobserved d-dimensional space. In the passive learning setting, assuming that the comparisons are samples from an unknown distribution over the underlying ordering, Ammar and Shah [19] propose a maximum entropy method with $\Omega(N^2)$ pairwise comparisons. Under the same non-parametric model, Negahban et al. [20] learn the ordering via an iterative rank aggregation algorithm requiring a total of $\Omega(N \log N)$ comparisons in which each pair needs to be repeated $\Omega(\log N)$ times. Shah et al. [21] show that a minimax optimal estimator can estimate the preference matrix with $O(\log^2 N/N)$ error. By showing that the preference matrix has rank $r \ll N$ under a suitable transformation, Rajkumar and Agarwal [22] show that $\Omega(rN \log N)$ comparisons suffice.

Among parametric models, Hajek et al. [23] show that the maximum likelihood estimator under Plackett-Luce model Plackett [24] requires $\Omega(N \log N)$ comparisons to learn Plackett-Luce scores. Vojnovic and Yun [25] show that estimating Thurstone [14] scores via MLE requires $O(N \log N/\lambda)$ comparisons, where λ is the smallest nonzero eigenvalue of the Laplacian of a graph generated by comparisons. Assuming comparison labels are independent, Ailon [26] proposes an active learning algorithm that requires $\Omega(\epsilon^{-6} N \log^5 N)$ comparison labels for a risk of ϵ times the optimal risk, where risk is a function that is minimized at the correct ordering. Spectral ranking methods also learn sample scores with theoretical guarantees. Negahban et al. [27] show that the rank centrality algorithm learns scores in $\Theta(N \log^3 N)$ comparisons, while several algorithms generalize this setting and improve upon this bound [28, 29]. For example, ASR [29] learns scores in $\Omega(\xi^{-2} m^3 N \text{poly}(\log N))$ m-way comparisons, error on BTL parameters where ξ is the spectral gap of the graph Laplacian.

The *rank regression* setting we study departs from the above works in regressing rankings from sample features, as opposed to comparisons alone. Even though inference algorithms for ranking regression and applications abound [6, 7], in contrast to rank aggregation, sample complexity results are sparse. Using independent pairwise comparisons, Canonne et al. [30] propose an algorithm over sample pairs that tests whether the empirical distribution is close to a target distribution. Kane et al. [31] propose an active learning algorithm to infer class labels via a special pairwise comparison oracle that indicates which sample is closer to the separating hyperplane of class labels. Both of these settings significantly depart from the one we consider here.

Our model encompasses the Bradley-Terry [13] and Thurstone [14] models; under both, our setting can be seen as learning a linear classifier over sample differences. Learning linear classifiers is of course classic in both the standard PAC learning setting [32, 33] and variants, including agnostic [34, 35] and active [36, 37] learning. We stress that all of the above works operate on linear classifiers under the assumption of i.i.d. samples, and therefore *do not readily generalize or apply to our setting*. This is precisely because pairs of samples are correlated, a phenomenon that is not present in standard PAC learning.

Closer to us, Niranjan and Rajkumar [38] and Chiang et al. [39] analyze pairwise rank regression. Niranjan and Rajkumar [38] recover the correct ranking with $N = \Omega(d^2)$ samples and a number of comparisons that are polylogarithmic in N, while Chiang et al. [39] provide a guarantee that depends on the ℓ_2-distortion (due to noise) of the pairwise comparison matrix. Nevertheless, both works ignore dependence across sample pairs. In particular, they analyze the concentration of labels over pairs of samples using Rademacher complexity bounds from Bartlett and Mendelson [40], that apply only if sample pair differences $\boldsymbol{x}_i - \boldsymbol{x}_j$ are independent. As a result, guarantees provided in [38] and [39] hold if every sample appears in only a single pair, an assumption we do not make. We further elaborate on this issue in Section 3.

3. Problem Formulation

Notation. For $N \in \mathbb{N}$, we denote by $[N] \equiv \{1, 2, \ldots, N\} \subset \mathbb{N}$ the set of integers from 1 to N, and use $\|\cdot\|$ for Euclidean (spectral) norm of vectors (matrices). The minimum and maximum singular values of a matrix $\boldsymbol{A}$ is denoted with $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$, respectively. We denote by $\mathbb{1}_{\mathcal{A}}$ the indicator function of a predicate $\mathcal{A}$, i.e., $\mathbb{1}_{\mathcal{A}} = 1$ if $\mathcal{A}$ is true and 0 otherwise.

Generative Model. We consider a setting in which an expert is presented with pairs of samples from a dataset. The expert produces a (possibly noisy) *comparison label* for each pair, *i.e.*, the expert selects among two samples the one ranked higher with respect to an underlying total ordering of the samples. Formally, we are given a dataset of $2N$ samples, each denoted by $i \in [2N]$. Each sample i has a corresponding feature vector $\boldsymbol{X}_i \in \mathbb{R}^d$. Using the first half of the dataset (i.e., $[N]$), the expert is presented with M

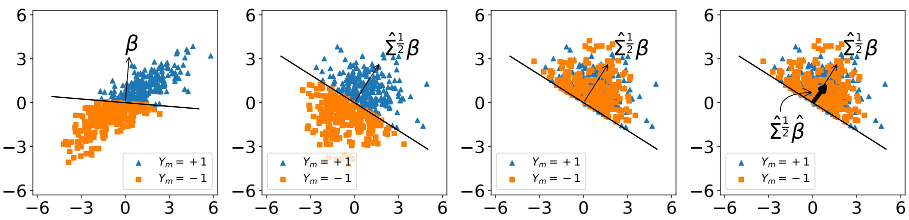


Figure 1: Intuition behind the estimator in Eq. (3). We consider a dataset of i.i.d. Gaussian samples $\{\boldsymbol{X}_i\}_{i=1}^{2N}$. Differences $\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m}$ are shown in Fig. 1a, along with $\boldsymbol{\beta}$ and the corresponding separating hyperplane. Colors indicate labels $Y_m \in \{-1,+1\}$. We can rewrite Eq. (3) as $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \cdot \frac{1}{M}\sum_{m=1}^{M} Y_m \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})$. Multiplying vectors $\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m}$ with $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ gives the whitened sample pairs in Fig. 1b; in this coordinate system, the separating hyperplane has normal $\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{\beta}$. The resulting whitened pairs are multiplied by the labels Y_m in Fig. 1c; this results in a "mirroring" over the separating hyperplane defined by $\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{\beta}$. Their average (i.e., $\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\hat{\boldsymbol{\beta}}$) is approximately co-linear with $\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{\beta}$. The final multiplication with $\hat{\boldsymbol{\Sigma}}^{-1/2}$ recovers $\boldsymbol{\beta}$ (up to a multiplicative constant).

pairs of samples $(I_m, J_m) \in [N] \times [N]$ where $m \in [M]$ and produces a comparison label $Y_m \in \{+1,-1\}$ where $Y_m = +1$ if I_m ranks higher than J_m and -1 otherwise. We denote the dataset of all comparisons by $\mathcal{D} = \{(I_m, J_m, Y_m)\}_{m=1}^{M}$.

We assume that the feature vectors $\boldsymbol{X}_i \in \mathbb{R}^d$ are independent and identically distributed (i.i.d.) Gaussian vectors with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive definite covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, i.e., $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We assume that the eigenvalues of $\boldsymbol{\Sigma}$ are ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$. Furthermore, we assume that I_m, J_m are sampled uniformly at random from $[N]$ and are independent of each other and $\{\boldsymbol{X}_i\}_{i=1}^{2N}$. Labels Y_m are independent of all other variables conditioned on $I_m, J_m, \boldsymbol{X}_{I_m}, \boldsymbol{X}_{J_m}$ and are distributed according to the following model: there exists a $\boldsymbol{\beta} \in \mathbb{R}^d$ such that the conditional distribution of Y_m is given by

$$\Pr(Y_m = 1 | \boldsymbol{X}_{I_m} = \boldsymbol{x}, \boldsymbol{X}_{J_m} = \boldsymbol{y}) = f(\boldsymbol{\beta}^\top(\boldsymbol{x} - \boldsymbol{y})), \tag{1}$$

where the function $f : \mathbb{R} \to [0,1]$ is (a) non-decreasing and continuously differentiable, and (b) satisfies

$$\lim_{x\to\infty} f(x) = 1, \quad \lim_{x\to-\infty} f(x) = 0, \quad f(-x) = 1 - f(x). \tag{2}$$

For example, $f(x)$ could be the sigmoid function, i.e. $f(x) = 1/(1+e^{-x})$, which results in the well known Bradley-Terry model [13]. Alternatively, f could be the cumulative distribution function of standard normal distribution, i.e. $f(x) = (1 + \text{erf}(x))/2$, which corresponds to the Thurstone model [14]. Both of these examples satisfy the aforementioned properties.

Parameter Estimation. The learner observes $\mathcal{D}$ and estimates $\boldsymbol{\beta}$ via:

$$\hat{\boldsymbol{\beta}} = \tfrac{1}{M}\textstyle\sum_{m=1}^{M} Y_m \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m}) \in \mathbb{R}^d, \tag{3}$$

where $\hat{\boldsymbol{\Sigma}}$ is an estimator of $\boldsymbol{\Sigma}$, computed over the second half of the samples through:

$$\hat{\boldsymbol{\Sigma}} = \tfrac{1}{N-d-2}\textstyle\sum_{i=N+1}^{2N}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\top, \text{ where } \hat{\boldsymbol{\mu}} = \tfrac{1}{N}\textstyle\sum_{i=N+1}^{2N}\boldsymbol{X}_i. \tag{4}$$

Table 1: Summary of Notation

N	number of samples	$\boldsymbol{X}_i$	Gaussian feature vector
M	number of comparisons	d	dimensionality of a feature vector
$\\|\cdot\\|$	ℓ_2 (spectral) norm of vectors (matrices)	i, n	sample index in $[N]$
Y_m	comparison label	m	comparison index in $[M]$
I_m, J_m	uniform random variables in $[N]$	$\mathcal{D}$	comparison dataset
$[N]$	set of integers from 1 to N	$\boldsymbol{\beta}$	parameter vector/model in $\mathbb{R}^d$
c_i	constants		

Note that $\mathbb{E}[\hat{\boldsymbol{\Sigma}}^{-1}] = \boldsymbol{\Sigma}^{-1}$ (see, e.g., [41]). We separate the dataset in two halves to ensure the independence of $\hat{\boldsymbol{\Sigma}}$ from labels in $\mathcal{D}$. Eq. (3) resembles a two-class linear discriminant analysis (LDA) estimator (see, e.g., [42]) and is indeed unbiased up to a positive multiplicative constant (see Lemma 7); this is a consequence of Stein's Lemma [43], stated formally in Section 4. Fig. 1 provides some intuition as to why this is the case. Despite the simplicity of our proposed estimator, characterizing its sampling complexity poses a significant challenge. Non-asymptotic bounds establishing consistency typically rely on i.i.d. assumptions; this is indeed natural to assume for samples $\{\boldsymbol{X}_i\}_{i=1}^{2N}$. However, pairwise comparisons introduce correlations in labels $\{\boldsymbol{Y}_m\}_{m=1}^{M}$: this is precisely because samples are re-used in pairs. We stress that conditioning on $\{\boldsymbol{X}_i\}_{i=1}^{2N}$ *does not* resolve this issue, as labels are still dependent through random variables (I_m, J_m).

Discussion on Previous Bounds. The previously mentioned works [38, 39] provide sample complexity bounds for a similar setting. In Niranjan and Rajkumar [38], a class of generative models, namely *feature low rank* models, are considered. In their analysis, each comparison is required to be repeated K times. They assess their method w.r.t. the normalized Kendall's Tau error metric [38, 39]; to obtain an error less than ϵ, they require $M = \Omega(d^2 \log N/\epsilon^2)$ different comparisons and $K = \Omega(d^2 \log^2 N/\epsilon N^2)$ independent repetitions of each comparison. Most importantly, our method does not enforce repetitions of each comparison. In contrast, our analysis captures the same complexity for M, *i.e.*, we require $M = \Omega(d^2 \log^3 N/\epsilon^4)$; note that our accuracy ϵ is on the Euclidean norm for the unbiased parameter estimate. Moreover, the risk bound they utilize (see Theorem 8 in Bartlett and Mendelson [40]) in the proof of their Theorem 1 requires independent samples for inputs $\boldsymbol{X}$ and $\boldsymbol{Y}$. This property indeed does not hold, since in two comparisons that share the same item, inputs are correlated due to the same feature vector appearing twice. The same theorem is used in Chiang et al. [39] (see Lemma 1 therein), where a sample complexity of $O(\|r\|^2/\epsilon^2)$, where $\|r\|^2$ is $o(N)$. Because the same theorem is used, and the theorem requires independent samples, the same issue arises. Our analysis takes into account these correlations and therefore corresponds to a more realistic setting. Our contribution is theoretical in nature; the advantage of our algorithm is that it is amenable to an analysis that handles this dependence. Nevertheless, we compare our method empirically with the method in Chiang et al. [39] in Section 7.

4. Technical Preliminary

In this section, we review some known results. The first is a variant of Stein's lemma from Liu [44]; we use this to show that our estimator is unbiased up to a constant.

Lemma 1 (Stein's Lemma [43, 44]). *Let $\boldsymbol{X} \in \mathbb{R}^d$, $\boldsymbol{X}' \in \mathbb{R}^{d'}$ be jointly Gaussian random vectors. Let the function $h : \mathbb{R}^{d'} \to \mathbb{R}$ be differentiable almost everywhere and satisfy $\mathbb{E}\left[\left|\partial h\left(\boldsymbol{X}'\right)/\partial X_i\right|\right] < \infty$, $i \in [d']$, then* $\mathrm{Cov}\left(\boldsymbol{X}, h(\boldsymbol{X}')\right) = \mathrm{Cov}\left(\boldsymbol{X}, \boldsymbol{X}'\right)\mathbb{E}\left[\nabla h\left(\boldsymbol{X}'\right)\right]$.

The second lemma we utilize bounds the tail of the norm of standard Gaussian vectors.

Lemma 2 (Centralized Chi-Squared Tail Bound [45]). *Let $F_X(x;k)$ be the CDF of centralized chi-square distribution with k degrees of freedom. Then, $1 - F_X(zk;k) \leq (ze^{1-z})^{k/2}$ for $z > 1$.*

A consequence of the way we select random pairs is that the joint distribution of the number of times each sample is selected is multinomial. The next inequality provides a bound for such variables:

Lemma 3 (Bretagnolle-Huber-Carolle Inequality [46]). *Let $\{M_i\}_{i=1}^N$ be multinomial r.v.s with parameters M , $\{p_i\}_{i=1}^N$. Then* $\Pr\left(\sum_{i=1}^N \left|\frac{M_i}{M} - p_i\right| \geq \epsilon\right) \leq 2^N e^{-\frac{\epsilon^2 M}{2}}$.

We also state the following classic inequality:

Lemma 4 (Hoeffding's Inequality [47]). *Let $X = \frac{1}{N}\sum_{i=1}^N X_i$, where $a_i \leq X_i \leq b_i$ and X_i are independent, and $\mu = \mathbb{E}[X]$. Then* $\Pr(|X - \mu| \geq \epsilon) \leq 2e^{-2N^2\epsilon^2/\sum_{i=1}^N (b_i - a_i)^2}$.

Recall that a random variable $X \in \mathbb{R}$ is sub-gaussian if there exists a $c > 0$ for all $t \geq 0$ s.t. $\Pr(|X| > t) \leq 2\exp(-t^2/c)$. Then, we define the sub-gaussian norm of X, denoted by $\|X\|_{\psi_2}$ as $\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[e^{X^2/t^2}\right] \leq 2\right\}$. Moreover, a random vector $\boldsymbol{X} \in \mathbb{R}^d$ is called sub-gaussian if one dimensional marginals $\boldsymbol{v}^\top \boldsymbol{X}$ are sub-gaussian for all $\boldsymbol{v} \in S^{d-1}$, where $S^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$. The sub-gaussian norm of $\boldsymbol{X}$ is then defined as $\|\boldsymbol{X}\|_{\psi_2} = \sup_{\boldsymbol{v} \in S^{d-1}} \left\|\boldsymbol{v}^\top \boldsymbol{X}\right\|_{\psi_2}$. The next lemma provides lower and upper bounds for the singular values of random design matrices.

Lemma 5 (Theorem 5.39 of Vershynin [48]). *Let $\boldsymbol{A} \in \mathbb{R}^{N \times d}$ be a matrix whose rows $\boldsymbol{A}_i$ are independent sub-gaussian isotropic random vectors. Then for every $t \geq 0$, with probability at least $1 - 2e^{-ct^2}$ one has $\sqrt{N} - C\sqrt{d} - t \leq \lambda_{\min}[\boldsymbol{A}] \leq \lambda_{\max}[\boldsymbol{A}] \leq \sqrt{N} + C\sqrt{d} + t$ where $c, C > 0$ depend only on $\max_i \|\boldsymbol{A}_i\|_{\psi_2}$.*

We use Lemma 5 to bound the eigenvalues of the feature covariance matrix. Lastly, the next lemma is used for bounding the norm of sub-gaussian random vectors.

Lemma 6 (Theorem 1 of Hsu et al. [49]). *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be a matrix, and let $\Sigma \equiv \boldsymbol{A}^\top \boldsymbol{A}$. Suppose that $\boldsymbol{x} \in \mathbb{R}^d$ is a sub-gaussian random vector with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\sigma = \|\boldsymbol{x}\|_{\psi_2}$. For all $t > 0$,* $\Pr(\|\boldsymbol{A}\boldsymbol{x}\|^2 > \sigma^2(\mathrm{Tr}(\Sigma) + 2\sqrt{\mathrm{Tr}(\Sigma^2)t} + 2\|\Sigma\| t) + \|\boldsymbol{A}\boldsymbol{\mu}\|^2 (1 + 4(\frac{\|\Sigma\|^2}{\mathrm{Tr}(\Sigma^2)}t)^{1/2} + \frac{4\|\Sigma\|^2}{\mathrm{Tr}(\Sigma^2)}t)^{1/2}) \leq e^{-t}$.

5. Main Results

We first establish that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ up to a multiplicative constant.

Lemma 7. *For $\hat{\boldsymbol{\beta}}$ in Eq.* (3), $\mathbb{E}[\hat{\boldsymbol{\beta}}]=c_1\boldsymbol{\beta}$, *where* $c_1=4\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T(\boldsymbol{X}_{I_m}-\boldsymbol{X}_{J_m})\right)\right]>0$.

The proof can be found in Appendix A. This result is a consequence of Stein's lemma [43] (see Lemma 1 in Section 4). Learning $\boldsymbol{\beta}$ up to a multiplicative constant suffices, as only the direction is enough to reveal the separating hyperplane between positive and negative sample pairs. Constant c_1 captures label noise: by (2), f' is non-negative and maximized at zero; for functions f that are "flatter" around zero the maximum value of f' and, therefore, c_1 is smaller. This is expected, as such f also result in noisier labels. Crucially, although our guarantees depend on c_1 (see Theorem. 1 below), our estimator does not depend on c_1: no knowledge of c_1 is required to compute $\hat{\boldsymbol{\beta}}$ via Eq. (3). Theorem 1 establishes that the parameters $\boldsymbol{\beta}$ are PAC learnable.

Theorem 1. *For $0<\epsilon<1$, sample count $N/\log^2 N=\Omega\left(\frac{d}{\epsilon^2\lambda_d}\right)$ and comparison count $M=\Omega\left(\frac{dN\log^3 N}{\epsilon^2\lambda_d}\right)$,*

$$\Pr\left(\left\|\hat{\boldsymbol{\beta}}-c_1\boldsymbol{\beta}\right\|\geq\epsilon\right)\leq c_2 N\max\left\{\left(\frac{\sqrt{6\log N}}{N}\right)^d, e^{-\frac{N\epsilon^2\lambda_d}{c_3 d\log N}}\right\}, \tag{5}$$

where $c_1>0$ is given by Lemma 7 and $c_2,c_3>0$ are absolute constants.

Theorem 1, which we prove below, allows us to characterize the sample complexity of $\hat{\boldsymbol{\beta}}$ in terms of the ambient dimension d, number of samples N, and number of comparisons M. It implies that to attain an accuracy $\epsilon>0$ with high probability, the estimator requires $\Omega(d\log^2 N/\epsilon^2\lambda_d)$ samples; this is of the same order as standard PAC learning guarantees for linear classifiers [50, 51, 52] and is also corroborated by our experiments in Section 7. Moreover, the number of comparisons required to attain an accuracy $\epsilon>0$ is $\Omega(dN\log^3 N/\epsilon^2\lambda_d)$, i.e., comparisons scale almost linearly with N.

We emphasize that, to identify the separating plane, it suffices to know $\boldsymbol{\beta}$ up to a non-negative multiplicative constant. This motivates the l.h.s. of Eq. (5) in Theorem 1. Nevertheless, the above guarantee should become more stringent for smaller $c_1>0$. Recalling that c_1 captures the level of label noise (smaller indicates more noise), the latter's impact on this bound is captured by replacing desired accuracy ϵ with $\epsilon'=c_1\epsilon$, so that c_1 appears as an additional constant in the r.h.s. of Eq. (5).

6. Proof of Theorem 1

The proof proceeds in the following manner. We first use a union bound to bound the tail of $\left\|\hat{\boldsymbol{\beta}}-c_1\boldsymbol{\beta}\right\|$ via several constituent terms. Contrary to standard concentration proofs, however, sums appearing in these terms involve dependent random variables. We nevertheless bound these terms by union bounds, conditioning, and leveraging the boundedness of random variables summed. From a technical standpoint, we leverage the Bretagnolle-Huber-Carolle inequality (see Lemma 3), and combine it with classic concentration inequalities (like Hoeffding's inequality, Lemma 4, and Lemma 16, due to [48]). We start with a simple bound on $\left\|\hat{\boldsymbol{\beta}}-c_1\boldsymbol{\beta}\right\|$.

Lemma 8. *For $\epsilon < 1$, the estimator $\hat{\boldsymbol{\beta}}$ given by Eq.* (3) *satisfies:*

$$\Pr\left(\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| > \epsilon\right) \leq 4\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\right\| \cdot \|\mathbb{E}[Y_m(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\| > \epsilon/6\right)$$
$$+ 4\Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\| > \frac{\sqrt{\lambda_d}\epsilon}{6}\right).$$

The proof, via a union bound, can be found in Appendix B. The terms Y_m, $\boldsymbol{X}_{I_m} - \boldsymbol{\mu}$ are not independent. This is because (a) the same sample $\boldsymbol{X}_i$ can be selected more than once, and, crucially, (b) the labels Y_m are coupled via the selection of the second sample in each pair. As a consequence, standard concentration bounds do not immediately apply. As a remedy, we condition on events under which the above variables are independent and refine this bound further. To do so, we introduce several quantities of interest. Let

$$\boldsymbol{W}_n = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_n - \boldsymbol{\mu}), \tag{6}$$

be the normalized feature vectors. For $n \in [N]$, let the number of times $I_m = n$ be

$$M_n = \textstyle\sum_{m=1}^{M} \mathbb{1}_{I_m = n}. \tag{7}$$

For $n, j \in [N]$, let $g_{n,j} : (\mathbb{R}^d)^2 \to [-1, 1]$ be the expected comparison label conditioned on the features of samples $n, j \in [N]$ selected in a pair, i.e.:

$$g_{n,j}(\boldsymbol{x}_n, \boldsymbol{x}_j) = \mathbb{E}[Y_m | I_m = n, J_m = j, \{\boldsymbol{X}_{n'} = \boldsymbol{x}_{n'}\}_{n'=1}^{N}] = 2f(\boldsymbol{\beta}^\top(\boldsymbol{x}_n - \boldsymbol{x}_j)) - 1.$$

Let $g_n : \mathbb{R}^d \to [-1, 1]$ be the expected label conditioned on the I_m-th sample:

$$g_n(\boldsymbol{x}) = \mathbb{E}[Y_m \mid I_m = n, \boldsymbol{X}_n = \boldsymbol{x}] = \textstyle\int g_{n,1}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{f}_{\boldsymbol{X}_1}(\boldsymbol{y}) \boldsymbol{dy}. \tag{8}$$

We will also need a similar quantity, $\tilde{g}_n : (\mathbb{R}^d)^N \to [-1, 1]$:

$$\tilde{g}_n(\{\boldsymbol{x}_{n'}\}_{n'=1}^{N}) = \mathbb{E}[Y_m \mid I_m = n, \{\boldsymbol{X}_{n'} = \boldsymbol{x}_{n'}\}_{n'=1}^{N}] = \tfrac{1}{N}\textstyle\sum_{j=1}^{N} g_{n,j}(\boldsymbol{x}_n, \boldsymbol{x}_j). \tag{9}$$

Note that g_n and $\tilde{g}_n$ are distinct, but the latter can be seen as a quantity that concentrates to g_n as N becomes large. We denote by $z_n : (\mathbb{R}^d)^N \to [-2, 2]$ their difference, i.e.:

$$z_n(\{\boldsymbol{x}_{n'}\}_{n'=1}^{N}) = \tilde{g}_n(\{\boldsymbol{x}_{n'}\}_{n'=1}^{N}) - g_n(\boldsymbol{x}_n). \tag{10}$$

Finally, let $\Delta_n : (\mathbb{R}^d)^N \to [-2, 2]$ be the difference between true label averages and $\tilde{g}_n$:

$$\Delta_n(\{\boldsymbol{x}_{n'}\}_{n'=1}^{N}) = \tfrac{1}{M_n}\textstyle\sum_{m:\, I_m = n} Y_m - \tilde{g}_n(\{\boldsymbol{x}_{n'}\}_{n'=1}^{N}). \tag{11}$$

Our next lemma bounds the second term in the r.h.s. of Lemma 8.

Lemma 9. *For $M_n, g_n, \tilde{g}_n, z_n, \Delta_n$ given by Equations* (7), (8), (9), (10), (11),

$$\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\|$$

$$\leq \left\| \sum_{n=1}^{N} \left(\frac{M_n}{M} - \frac{1}{N} \right) \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| + \left\| \frac{1}{M} \sum_{n=1}^{N} \boldsymbol{W}_n M_n \Delta_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\|$$
$$+ \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n z_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| + \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n g_n(\boldsymbol{X}_n) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})] \right\|.$$

The proof can be found in Appendix C. The four terms in the r.h.s. are bounded individually in the rest of the proof. We bound the first term in Lemma 9 with Lemma 10.

Lemma 10. *For all $\delta_0 > d$,*

$$\Pr\left(\left\| \sum_{n=1}^{N} \left(\frac{M_n}{M} - \frac{1}{N} \right) \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| > \epsilon \right) \leq N \left(\frac{\delta_0}{d} e^{1 - \frac{\delta_0}{d}} \right)^{d/2} + 2^N e^{-\frac{\epsilon^2 M}{2\delta_0}}.$$

The proof can be found in Appendix D. We rely on the fact that $|\tilde{g}_n| \leq 1$, as well as (a) the norm $\|\boldsymbol{W}_n\|$ can be bounded by a centralized Chi-Squared tail bound, while (b) the quantity $|\frac{M_n}{M} - \frac{1}{N}|$ can be bounded by the Bretagnolle-Huber-Carol Inequality (see Lemma 3 in Section 4). Next, we bound the second term in Lemma 9.

Lemma 11. *For all $\delta_1 < \frac{\epsilon^2}{4d}$ and $\delta_2 > d$,*

$$\Pr\left(\left\| \frac{1}{M} \sum_{n=1}^{N} \boldsymbol{W}_n M_n \Delta_n \{\boldsymbol{X}_{n'}\}_{n'=1}^{N} \right\| > \epsilon \right) \leq N \left(\frac{\epsilon^2}{4d\delta_1} e^{1 - \frac{\epsilon^2}{4d\delta_1}} \right)^{d/2}$$
$$+ N \left(\frac{\delta_2}{d} e^{1 - \frac{\delta_2}{d}} \right)^{d/2} + 2^N e^{-\frac{\epsilon^2 M}{8\delta_1\delta_2}} + 2e^{\log N - \frac{M\delta_1}{2N} - o(\frac{M\delta_1}{2N})}.$$

The proof is in Appendix E. We bound individual terms, $\|\boldsymbol{W}_n\|$, $|\Delta_n|$, $\left|\frac{M_n}{M} - \frac{1}{N}\right|$ respectively using a centralized Chi-Squared tail bound, Hoeffding's inequality, and the moment generating function of the binomial distribution. Our next lemma bounds the third term in Lemma 9.

Lemma 12. *For all $\delta_3 \leq \epsilon^2/d$,*

$$\Pr\left(\left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n z_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| > \epsilon \right) \leq N \left(\frac{\epsilon^2}{d\delta_3} e^{1 - \frac{\epsilon^2}{d\delta_3}} \right)^{d/2} + 2Ne^{-\frac{N\delta_3}{2}}.$$

The proof can be found in Appendix F. We bound terms $\|\boldsymbol{W}_n\|$ and $|z_n|$ individually. For the former, we again use a centralized Chi-Squared tail bound. For the latter, we indeed show that, for large sample sizes N, $\tilde{g}_n$ concentrates around g_n using Hoeffding's inequality. We bound the last term in Lemma 9 as follows

Lemma 13. *For an absolute constant $c_2 > 0$,*

$$\Pr\left(\left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n g_n(\boldsymbol{X}_n) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})] \right\| > \epsilon \right) \leq e^{-\frac{1}{4}\left(\sqrt{\frac{N\epsilon^2}{c_2} - d} - \sqrt{d}\right)^2}.$$

The proof, which is in Appendix G, shows that individual terms are sub-gaussian and uses a concentration bound due to Hsu et al. [49]. The second term in Lemma 8 is bounded as follows.

Lemma 14. *For the estimator $\hat{\boldsymbol{\Sigma}}$ given by Eq.* (4)*, and for* $\sqrt{N} > \sqrt{\frac{N-d-2}{\frac{\lambda_d}{d\sqrt{2\lambda_1}}\epsilon+1}} + c_3\sqrt{d}$ *where* $c_3, c_4 > 0$ *are absolute constants,*

$$\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\right\| \cdot \|\mathbb{E}[Y_m(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\| > \epsilon\right) \leq 2e^{-c_4\left(\sqrt{N} - \sqrt{\frac{N-d-2}{\frac{\lambda_d}{d\sqrt{2\lambda_1}}\epsilon+1}} - c_3\sqrt{d}\right)^2}.$$

The proof can be found in Appendix H. We use a concentration bound on the minimum singular value of the design matrix due to Vershynin [48]. Combining Lemmas 8, 9, 10, 11, 12, 13 and 14 via the union bound gives:

$$\begin{aligned}
\Pr\left(\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| \geq \epsilon\right) \leq\ & 8e^{-c_4\left(\sqrt{N} - \sqrt{\frac{N-d-2}{\frac{\lambda_d}{6d\sqrt{2\lambda_1}}\epsilon+1}} - c_3\sqrt{d}\right)^2} + 4e^{-\frac{1}{4}\left(\sqrt{\frac{\epsilon^2 N\lambda_d}{c_2}} - d - \sqrt{d}\right)^2} \\
& + 2^{N+2}e^{-\frac{\epsilon^2 M\lambda_d}{4608\delta_1\delta_2}} + 8e^{\log N - \frac{M\delta_1}{2N} - o\left(\frac{M\delta_1}{2N}\right)} + 8Ne^{-\frac{N\delta_3}{2}} \\
& + 4N\left(\frac{\epsilon^2\lambda_d}{2304d\delta_1}e^{1-\frac{\epsilon^2\lambda_d}{2304d\delta_1}}\right)^{d/2} + 4N\left(\frac{\epsilon^2\lambda_d}{576d\delta_3}e^{1-\frac{\epsilon^2\lambda_d}{576d\delta_3}}\right)^{d/2} \\
& + 2^{N+2}e^{-\frac{\epsilon^2 M\lambda_d}{1152\delta_0}} + 4N\left(\frac{\delta_0}{d}e^{1-\frac{\delta_0}{d}}\right)^{d/2} + 4N\left(\frac{\delta_2}{d}e^{1-\frac{\delta_2}{d}}\right)^{d/2}.
\end{aligned}$$

Setting $M = \Omega(\frac{dN\log^3 N}{\lambda_d\epsilon^2})$, $\delta_0 = d\log^2 N$, $\delta_1 = 4\lambda_d\epsilon^2/d\log^2 N$, $\delta_2 = d\log^2 N$ and $\delta_3 = \epsilon^2\lambda_d/1152d\log N$, the bound reduces to $\Pr\left(\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| \geq \epsilon\right) \leq c_6 N \max\left\{\left(\frac{\sqrt{6\log N}}{N}\right)^d, e^{-\frac{N\epsilon^2\lambda_d}{c_7 d\log N}}\right\}$ for $N > \frac{c_8 d\log^2 N}{\epsilon^2\lambda_d}$ and $0 < \epsilon < 1$, where $c_6, c_7, c_8 > 0$ are absolute constants. We derive this in Appendix I. □

7. Experiments

Synthetic Experiment Setup. To support our theoretical findings, we evaluate (Code available online: `https://git.io/Jkbk1`) the estimator given by Eq. (3) with a synthetic dataset as follows: We sample the true parameter $\boldsymbol{\beta} \in \mathbb{R}^d$ from $\mathcal{N}(\mathbf{0}, 10\,\boldsymbol{I})$. We sample $\boldsymbol{\mu} \in \mathbb{R}^d$ uniformly at random from $[-5, 5]^d$. To assess the impact of the minimum eigenvalue of $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ on estimator accuracy, we generate $\boldsymbol{\Sigma}$ as follows. We generate a random orthonormal basis of $\mathbb{R}^d$, and choose a smallest eigenvalue $\lambda_d \in (0, 1]$. We then a construct $\boldsymbol{\Sigma}$ whose eigenvectors are the selected orthonormal basis, and d eigenvalues equidistributed in $[\lambda_d, 1]$. We treat λ_d as a tunable parameter. Each feature vector $\boldsymbol{x}_i \in \mathbb{R}^d, i \in [2N]$, is independently sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We sample pairs $(I_m, J_m), m \in [M]$, uniformly at random from $[N] \times [N]$. Noisy labels Y_m are sampled using Eq. (1) where $f(x) = (1 + e^{-\alpha x})^{-1}$ and $0 < \alpha < \infty$. By adjusting α, we choose the fraction $p_e \in [0, 1]$ of M comparisons that are flipped, *i.e.*

are incorrect. We repeat all experiments with 10 random generations of parameters $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$. In synthetic experiments, we know the data statistics and the true latent parameter, therefore we can compute $c_1 \in \mathbb{R}$ from Lemma 7 and the probability of a label being flipped $p_e \in [0,1]$ for a sigmoid with adjustable slope $f(x) = (1+e^{-\alpha x})^{-1}$ where $\alpha > 0$. We explain how to compute these values below.

Computing c_1. We remind the reader that $c_1 = 4\mathbb{E}[f'(\boldsymbol{\beta}^\top(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m}))$. Let the score of item $i \in [2N]$ be $s_i = \boldsymbol{\beta}^\top \boldsymbol{X}_i \in \mathbb{R}$ and $s_{i,j} = \boldsymbol{\beta}^\top \boldsymbol{X}_i - \boldsymbol{\beta}^\top \boldsymbol{X}_j \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = 2\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}$ since $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note that c_1 is the result of a sigmoid-Gaussian type integral, *i.e.* $c_1 = 4\int f'(s) f_{s_{i,j}}(s) d_s$ where $f_{s_{i,j}}(s)$ is the probability density function of the distribution of $s_{i,j}$. As this integral does not have an analytical solution, we compute it numerically using the trapezoidal rule by taking finely spaced values $s \in [-4\sigma, 4\sigma]$.

Computing p_e. We remind the reader that p_e is the probability of a true label being flipped and is a function of $f(x)$ in Eq. (1) and $\boldsymbol{\beta}, \boldsymbol{\Sigma}$. We show that

$$\begin{aligned}\Pr(\text{"Error in label } m\text{"}) &= \Pr(Y_m = 1 \cap s_i < s_j) + \Pr(Y_m = -1 \cap s_i > s_j) \\ &= \Pr(Y_m = 1 \cap s_{i,j} < 0) + \Pr(Y_m = -1 \cap s_{i,j} > 0).\end{aligned}$$

Note that,

$$\begin{aligned}\Pr(Y_m = 1 \cap s_{i,j} < 0) &= \int_{-\infty}^{\infty} \Pr(Y_m = 1 \cap s_{i,j} < 0 | s_{i,j} = s) f_{s_{i,j}}(s) ds \\ &= \int_{-\infty}^{0} \Pr(Y_m = 1 | s_{i,j} = s) f_{s_{i,j}}(s) ds \\ &= \int_{-\infty}^{0} f(s) f_{s_{i,j}}(s) ds,\end{aligned}$$

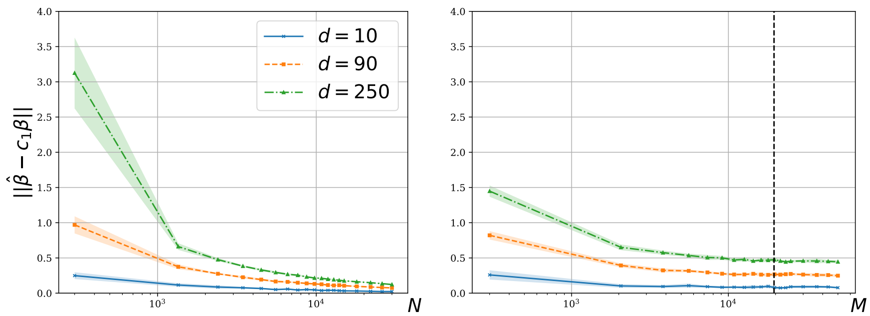
which is again a sigmoid-Gaussian type integral. Furthermore, it is straightforward to show that $\Pr(Y_m = 1 \cap s_{i,j} < 0) = \Pr(Y_m = -1 \cap s_{i,j} > 0)$. We evaluate this integral numerically using the trapezoidal rule by taking finely spaced values $s \in [-4\sigma, 0]$.

Metrics. We measure the performance of the estimator $\hat{\boldsymbol{\beta}}$ with two metrics. For consistency with our analysis, our first error metric is $\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\|$. The second metric is

$$\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \cos^{-1}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{\beta} / ||\hat{\boldsymbol{\beta}}|| ||\boldsymbol{\beta}||), \tag{12}$$

i.e., the angle between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. We can evaluate this without computing c_1; this makes the angle a more practical way to measure if the direction of $\boldsymbol{\beta}$ is correctly learned, that can also be used when c_1 is not known. We report both the average and standard deviation across different random generations.

Convergence. To investigate the rate of convergence of $\hat{\boldsymbol{\beta}}$, we vary the number of samples N in $[300, 3 \times 10^4]$, dimensionality $d \in \{10, 90, 250\}$ while we set $M = \lceil N \log N \rceil$, and $\lambda_d = 1$. We select α so that $p_e = 0.2$. In Fig. 2a, we plot the error $\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\|$ as a function of the dataset size N. We observe that for each d, the error decreases as N increases and $\hat{\boldsymbol{\beta}}$ indeed converges to $c_1\boldsymbol{\beta}$. In Fig. 2b, we vary M in



(a) Error by N

(b) Error by M

Figure 2: (a) The error of the estimator given by Eq. (3) indeed reduces as N increases when $M = \lceil N \log N \rceil$ and we see that the estimator is converging to $c_1\boldsymbol{\beta}$. (b) The error reduces when M increases while N is kept constant, however the decay is insignificant after $M = N \log N$, which is denoted with the black dashed line. This agrees with our theory that $M = \tilde{\Omega}(N)$. The shaded area is the standard deviation.

$[300, 5 \times 10^4]$ and $d \in \{10, 90, 250\}$ while we set $N = 2.5 \times 10^3$. We observe that increasing M reduces the error. However, the reduction in error is insignificant after $M = N \log N$, which is denoted with the black dashed line in Fig. 2b. This is consistent with the bound in Theorem 1, which anticipates that M is polylogarithmic in N. This result also shows that our Theorem 1 can possibly be improved, *i.e.*, the factors d and $\log^2 N$ can be reduced from M.

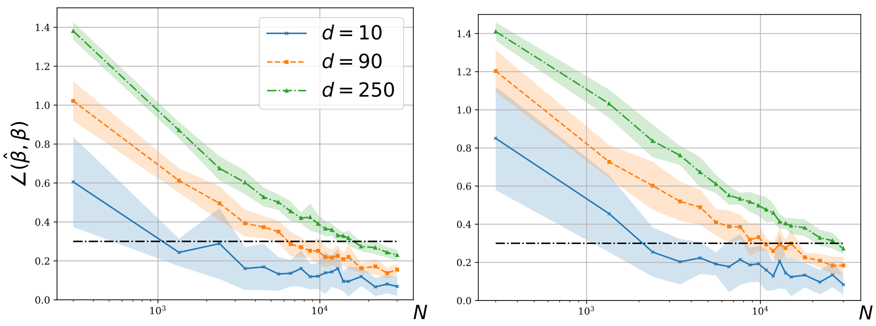
Dependence on N. We investigate the required number of samples N to attain $\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = 0.3$. For each $d \in \{10, 90, 250\}$ and $p_e \in \{0, 0.4\}$ we vary $N \in [300, 3 \times 10^4]$, setting $M = \lceil N \log N \rceil$, and $\lambda_d = 1/200$. In Fig. 3, we plot the error $\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ versus the dataset size N under different noise levels p_e. We observe that as N increases, $\hat{\boldsymbol{\beta}}$ indeed achieves $\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = 0.3$ for all d, while the error increase with d. This implies that irrespective of the noise level and the corresponding c_1 value, the estimator $\hat{\boldsymbol{\beta}}$ is able to recover the direction of $\boldsymbol{\beta}$ as N increases.

Dependence on M. We repeat our experiments on the impact of M, this time focusing on the angle metric and varying p_e. We fix $N = 2.5 \times 10^3$, $\lambda_d = 1/200$ and vary $M \in [300, 5 \times 10^4]$, $d \in \{10, 90, 250\}$, and $p_e \in \{0, 0.4\}$. Fig. 4 plots $\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ versus M under different noise levels. These plots show that the benefit of increasing M again diminishes beyond $M = N \log N$ for all $p_e \in \{0, 0.2, 0.4\}$, represented by the dashed black line. This is again consistent with Theorem 1.

Dependence on d. In Fig. 5, we plot the smallest N that achieves $\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq 0.3$ while we vary $d \in [10, 250]$, $\lambda_d \in \{1/200, 0.1, 1\}$ and $p_e \in \{0, 0.4\}$. We observe that the required N increases linearly in d. This is consistent with the linear dependence of N on d anticipated by Theorem 1.

Dependence on λ_d. In Fig. 6, we investigate the effect of λ_d, the smallest eigenvalue of $\boldsymbol{\Sigma}$, on the smallest N that achieves $\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq 0.3$ for each $p_e \in \{0, 0.4\}$ and $d \in \{10, 90, 250\}$. We observe the inversely proportional relationship between λ_d and N, which is consistent with the $N = \Omega(d/\lambda_d \epsilon^2)$ requirement implied by Theorem 1.

Evaluating Chiang et al. [39]. In this section, we evaluate the applicability of our



(a) $p_e = 0$

(b) $p_e = 0.4$

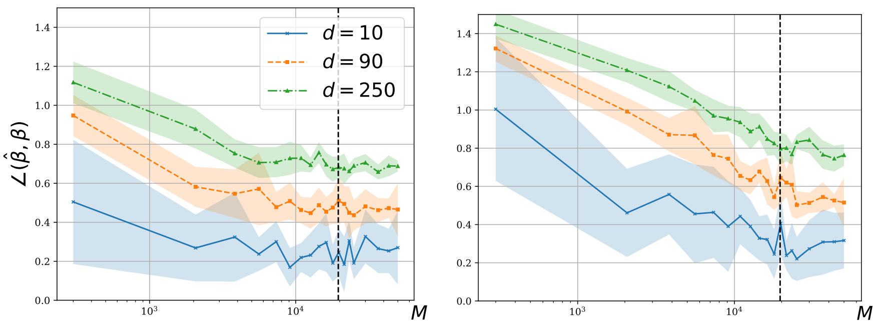
Figure 3: The angle between the true parameter $\boldsymbol{\beta}$ and the estimator $\hat{\boldsymbol{\beta}}$, plotted against N for different error probabilities p_e when $M = \lceil N \log N \rceil$ and $\lambda_d = 1/200$. (a) The noiseless case. (b) The case where 40% of the labels are flipped. Even though noise increases error, increasing N allows the estimator to reduce the error arbitrarily. This shows that the estimator $\hat{\boldsymbol{\beta}}$ is able to recover the direction of the true parameter $\boldsymbol{\beta}$ as N increases. The shaded area is the standard deviation.

Theorem 1 to the algorithm *RABF-log* defined in Chiang et al. [39]. We generate data the same as before with smaller values for N and d: We span $N \in [200, 2000]$ and $d \in \{10, 100\}$. In training *RABF-log*, we divide the dataset into train and validation splits to choose hyperparameters via grid search for 4 values that are logarithmically spaced in $[10^{-6}, 10^{3}]$. We also set α so that $p_e = 0.4$ and $\lambda_d = 1$.

Convergence of *RABF-log*. Fig. 7a shows that for *RABF-log*, the error reduces as N increases when $M = N \log N$. In Fig. 7b we keep $N = 1000$ and increase M. The black dashed line is the point where $M = N \log N$. We observe that for all d values, as M increases the error reduces. We also observe that for small d values, when M exceeds $N \log N$, the reduction is insignificant. However, for large d values, increasing M keeps reducing error. These are in line with our bound that requires $M = \Omega(dN \log^3 N/\epsilon^2)$.

Dependence of *RABF-log* to d and λ_d. In Fig. 8a, we plot the minimum N value that achieves angle 0.5 as d increases and we let $M = N \log N$. We observe that for different λ_d values, the curves are linear. This supports the linear dependence of N to d. In Fig. 8b, we plot plot the minimum N value that achieves angle 0.5 as λ_d increases when $M = N \log N$. We observe that for different d values, the curves depend on λ_d inversely proportionally; this is inline with our Theorem 1 that states $N = \Omega(d/\lambda_d)$.

Comparison to *RABF-log*. Empirically, the sampling complexity the *RABF-log* behaves similarly with our method as shown in Fig.s 7, 8. However, *RABF-log* is quite similar to an MLE solution for the parameters under our generative model, which is a convex optimization problem and *RABF-log* performs better in the chosen error metrics. This can be observed in Fig. 9a. One disadvantage of *RABF-log* is due to its optimization involving a parameter for each item in the dataset (N additional parameters compared to d alone). Therefore, it is significantly slower than our method; this can be observed in Fig. 9b.



(a) $p_e = 0$

(b) $p_e = 0.4$

Figure 4: The angle between the true parameter $\boldsymbol{\beta}$ and the estimator $\hat{\boldsymbol{\beta}}$, plotted against M for different error probabilities p_e when $N = 2.5 \times 10^3$ and $\lambda_d = 1/200$. (a) The noiseless case. (b) The case where 40% of the labels are flipped. Increasing M reduces the error for all p_e values. However, once $M = N \log N$ which is denoted with the black dashed line, the reduction is insignificant. This is due to the fact that a higher N is required for smaller ϵ as $N = \Omega(d/\lambda_d \epsilon^2)$, *i.e.* N needs to scale inversely quadratic with ϵ. This supports our theory that the estimator $\hat{\boldsymbol{\beta}}$ does better as M increases; however, for arbitrarily small ϵ, N needs to increase too. The shaded area is the standard deviation.

8. Conclusion

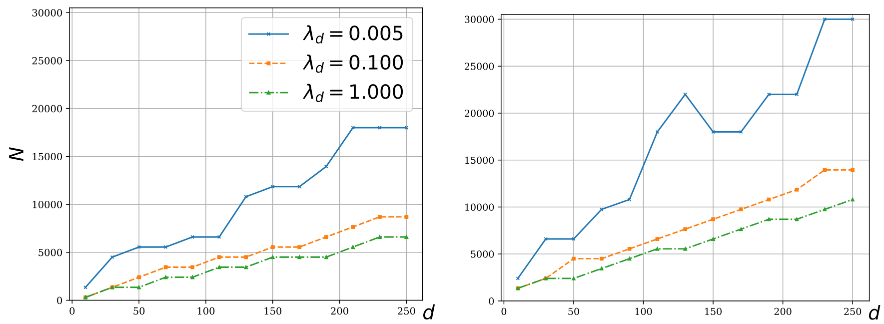
Our results suggest that learning parameters of a generalized linear preference model from comparisons come with guarantees, despite the lack of independence between comparisons. In particular, our results hold under u.a.r. comparisons and repetitions are permitted (due to sampling with replacement); our work can be used as a template to manage dependence of pairs in other learning settings where pairs are sampled with repetition. Though our bound on N is tight in terms of d (as the number of samples cannot be lower than d), our experimental results suggest that the bound on M could be sharpened w.r.t. d and N. Our approach serves as a starting point to the design and analysis of more sample efficient algorithms w.r.t. d as well as other parameters, such as the error ϵ or the smallest eigenvalue of the covariance λ_d. Finally, lower bounds on both N and M remain open, and would be an interesting future direction to investigate.

Acknowledgements

Our work is supported by NIH (R01EY019474), NSF (SCH-1622542 at MGH; SCH-1622536 at Northeastern; SCH-1622679 at OHSU), and by unrestricted departmental funding from Research to Prevent Blindness (OHSU).

References

[1] M. Schultz, T. Joachims, Learning a distance metric from relative comparisons, in: Advances in Neural Information Processing Systems (NeurIPS), 41–48, 2004.



(a) $p_e = 0$

(b) $p_e = 0.4$

Figure 5: Minimum N that achieves $\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq 0.3$ plotted against dimensionality d for different values of the minimum eigenvalue λ_d of the feature covariance and error probabilities p_e. (a) The noiseless case. (b) The case where 40% of the labels are flipped. As the probability of error and the condition number of the feature covariance increases, we require more samples. Crucially, we observe the linear dependence of N to d and this supports that $N = \Omega(d/\lambda_d \epsilon^2)$.

[2] Y. Zheng, L. Zhang, X. Xie, W. Y. Ma, Mining interesting locations and travel sequences from GPS trajectories, in: International Conference on World Wide Web (WWW), 791–800, 2009.

[3] Y. Koren, J. Sill, OrdRec: An ordinal model for predicting personalized item rating distributions, in: ACM Conference on Recommender Systems (RecSys), 117–124, 2011.

[4] D. McFadden, Conditional Logit Analysis of Qualitative Choice Behavior, Institute of Urban and Regional Development, University of California, 1973.

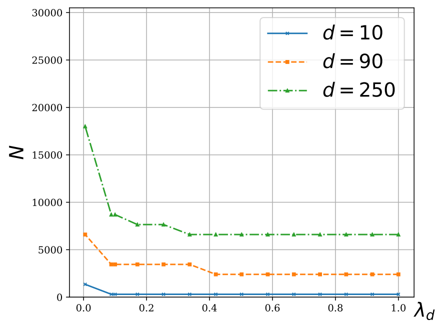
[5] G. van Ryzin, S. Mahajan, On the relationship between inventory costs and variety benefits in retail assortments, Management Science 45 (11) (1999) 1496–1509.

[6] P. Tian, Y. Guo, J. Kalpathy-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, J. Dy, D. Erdogmus, S. Ioannidis, A severity score for retinopathy of prematurity, in: International Conference on Knowledge Discovery & Data Mining (KDD), 1809–1819, 2019.

[7] İ. Yıldız, P. Tian, J. Dy, D. Erdoğmuş, J. Brown, J. Kalpathy-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, S. Ioannidis, Classification and comparison via neural networks, Neural Networks 118 (2019) 65–80.

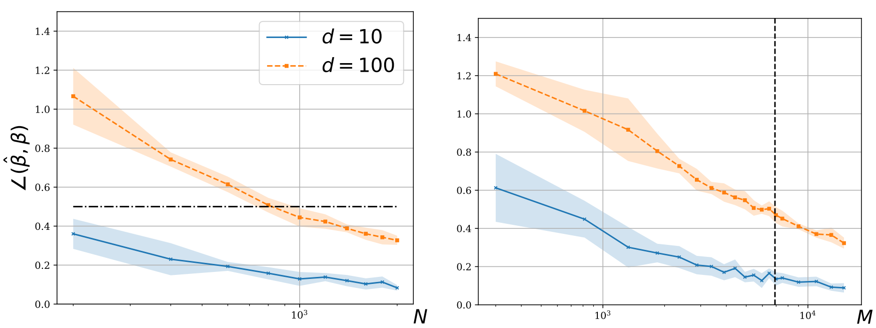
[8] Y. Guo, J. Dy, D. Erdoğmuş, J. Kalpathy-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, S. Ioannidis, Variational Inference from Ranked Samples with Features, in: Asian Conference on Machine Learning (ACML), 599–614, 2019.

[9] J. P. Campbell, J. Kalpathy-Cramer, D. Erdogmus, P. Tian, D. Kedarisetti, C. Moleta, J. D. Reynolds, K. Hutcheson, M. J. Shapiro, M. X. Repka, et al., Plus disease in retinopathy of prematurity: A continuous spectrum of vascular abnormality as a basis of diagnostic variability, Ophthalmology 123 (11) (2016) 2338–2344.



(a) $p_e = 0$ (b) $p_e = 0.4$

Figure 6: Minimum N that achieves $\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq 0.3$ plotted against the minimum eigenvalue λ_d of the feature covariance for different dimensionality d and probability of error p_e. (a) The noiseless case. (b) The case where 40% of the labels are flipped. We observe that increasing error increases the required N and the inversely proportional dependence of N to λ_d is observed.



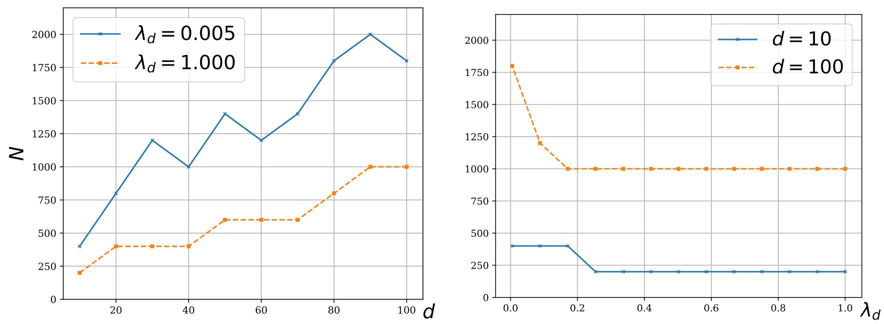
(a) Error by N (b) Error by M

Figure 7: The angle between the true parameter $\boldsymbol{\beta}$ and the $\hat{\boldsymbol{\beta}}$ from *RABF-log*, plotted against N and M for error probability $p_e = 0.4$ and $\lambda_d = 1$. (a) Increasing N when $M = \lceil N \log N \rceil$ reduces error arbitrarily. (b) Increasing M when N is constant also reduces error. Note that the dashed line is $M = N \log N$ and the reduction in error is insignificant after the dashed line for small d and significant when d is large. These are inline with our theoretical bound that $M = \Omega(dN \log^3 N/\epsilon^2)$.

[10] J. Kalpathy-Cramer, J. P. Campbell, D. Erdogmus, P. Tian, D. Kedarisetti, C. Moleta, J. D. Reynolds, K. Hutcheson, M. J. Shapiro, M. X. Repka, et al., Plus disease in retinopathy of prematurity: Improving diagnosis by ranking disease severity and using quantitative image analysis, Ophthalmology 123 (11) (2016) 2345–2351.

[11] N. Stewart, G. D. Brown, N. Chater, Absolute identification by relative judgment, Psychological Review 112 (4) (2005) 881–911.

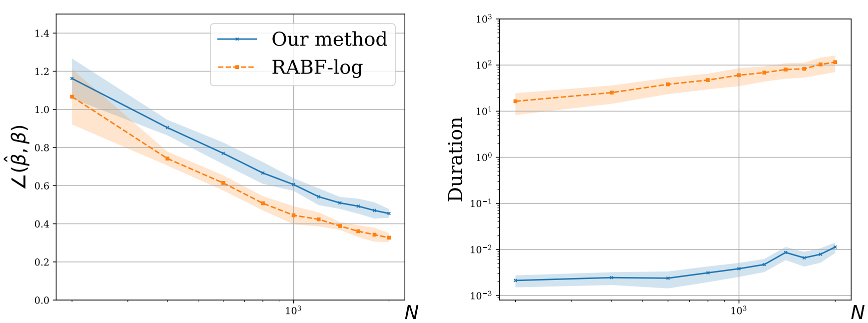
[12] S. Liu, T. Tran, G. Li, Y. Jiang, Ordinal random fields for recommender systems, in: Asian Conference on Machine Learning (ACML), 283–298, 2014.



(a) $p_e = 0.4$

(b) $p_e = 0.4$

Figure 8: Minimum N that achieves an accuracy ϵ for different settings. (a) For different λ_d values we observe that the relationship between N and d is linear. (b) For different d values, we observe that the relationship between N and λ_d is inversely proportional. These are both in line with our bounds for $N = \Omega(d/\lambda_d)$ and $M = \Omega(dN \log^3 N)$.



(a) Error by N.

(b) Training duration by N.

Figure 9: A comparison between our method and *RABF-log* when $\lambda_d = 1$ and $p_e = 0.4$. (a) We observe that *RABF-log* achieves smaller angle throughout, compared to our method. (b) We also observe that the training duration (in seconds) for *RABF-log* is orders of magnitude larger than our method.

[13] R. A. Bradley, M. E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons, Biometrika 39 (3/4) (1952) 324–345.

[14] L. L. Thurstone, A law of comparative judgment, Psychological Review 34 (4) (1927) 266–270.

[15] C. Dwork, R. Kumar, M. Naor, D. Sivakumar, Rank aggregation methods for the web, in: International Conference on World Wide Web (WWW), 613–622, 2001.

[16] Y. Zheng, Y. Zheng, D. Suehiro, S. Uchida, Top-rank convolutional neural network and its application to medical image-based diagnosis, Pattern Recognition 120 (2021) 108138.

[17] M. Braverman, E. Mossel, Noisy sorting without resampling, in: Symposium on Discrete Algorithms (SODA), 268–276, 2008.

[18] K. G. Jamieson, R. Nowak, Active ranking using pairwise comparisons, in: Advances in Neural Information Processing Systems (NeurIPS), 2240–2248, 2011.

[19] A. Ammar, D. Shah, Ranking: Compare, don't score, in: Allerton Conference on Communication, Control, and Computing (Allerton), 776–783, 2011.

[20] S. Negahban, S. Oh, D. Shah, Iterative ranking from pair-wise comparisons, in: Advances in Neural Information Processing Systems (NeurIPS), 2474–2482, 2012.

[21] N. Shah, S. Balakrishnan, A. Guntuboyina, M. Wainwright, Stochastically transitive models for pairwise comparisons: Statistical and computational issues, in: International Conference on Machine Learning (ICML), 11–20, 2016.

[22] A. Rajkumar, S. Agarwal, When can we rank well from comparisons of $O(n \log(n))$ non-actively chosen pairs?, in: Conference on Learning Theory (COLT), 1376–1401, 2016.

[23] B. Hajek, S. Oh, J. Xu, Minimax-optimal inference from partial rankings, in: Advances in Neural Information Processing Systems (NeurIPS), 1475–1483, 2014.

[24] R. L. Plackett, The analysis of permutations, Journal of the Royal Statistical Society: Series C (Applied Statistics) 24 (2) (1975) 193–202.

[25] M. Vojnovic, S. Yun, Parameter estimation for generalized Thurstone choice models, in: International Conference on Machine Learning (ICML), 498–506, 2016.

[26] N. Ailon, An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity, Journal of Machine Learning Research 13 (2012) 137–164.

[27] S. Negahban, S. Oh, D. Shah, Rank centrality: Ranking from pairwise comparisons, Operations Research 65 (1) (2017) 266–287.

[28] L. Maystre, M. Grossglauser, Fast and accurate inference of Plackett–Luce models, in: Advances in Neural Information Processing Systems (NeurIPS), 172–180, 2015.

[29] A. Agarwal, P. Patil, S. Agarwal, Accelerated spectral ranking, in: International Conference on Machine Learning (ICML), 70–79, 2018.

[30] C. L. Canonne, D. Ron, R. A. Servedio, Testing probability distributions using conditional samples, SIAM Journal on Computing 44 (3) (2015) 540–616.

[31] D. M. Kane, S. Lovett, S. Moran, J. Zhang, Active classification with comparison queries, in: Symposium on Foundations of Computer Science (FOCS), 355–366, 2017.

[32] V. N. Vapnik, A. Y. Chervonenkis, On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities, Theory of Probability & Its Applications 16 (2) (1971) 264–280.

[33] M. F. Balcan, H. Zhang, Sample and computationally efficient learning algorithms under s-concave distributions, in: Advances in Neural Information Processing Systems (NeurIPS), 4796–4805, 2017.

[34] A. T. Kalai, A. R. Klivans, Y. Mansour, R. A. Servedio, Agnostically learning halfspaces, SIAM Journal on Computing 37 (6) (2008) 1777–1805.

[35] E. Mammen, A. B. Tsybakov, Smooth discrimination analysis, Annals of Statistics 27 (6) (1999) 1808–1829.

[36] P. Awasthi, M. F. Balcan, N. Haghtalab, H. Zhang, Learning and 1-bit compressed sensing under asymmetric noise, in: Conference on Learning Theory (COLT), 152–192, 2016.

[37] C. Zhang, Efficient active learning of sparse halfspaces, in: Conference on Learning Theory (COLT), vol. 75, 1–26, 2018.

[38] U. Niranjan, A. Rajkumar, Inductive pairwise ranking: going beyond the $n\log(n)$ barrier, in: AAAI Conference on Artificial Intelligence (AAAI), 2436–2442, 2017.

[39] K.-Y. Chiang, C.-J. Hsieh, I. Dhillon, Rank aggregation and prediction with item features, in: International Conference on Artificial Intelligence and Statistics (AISTATS), 748–756, 2017.

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

[41] J. Hartlap, P. Simon, P. Schneider, Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix, Astronomy & Astrophysics 464 (1) (2007) 399–404.

[42] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, Springer Series in Statistics, 2001.

[43] C. M. Stein, Estimation of the mean of a multivariate distribution, The Annals of Statistics 9 (6) (1981) 1135–1151.

[44] J. S. Liu, Siegel's formula via Stein's identities, Statistics & Probability Letters 21 (3) (1994) 247–251.

[45] S. Dasgupta, A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss, Random Structures & Algorithms 22 (1) (2003) 60–65.

[46] J. Wellner, A. W. van der Vaart, Weak Convergence and Empirical Processes: with Applications to Statistics, Springer Science & Business Media, 2013.

[47] W. Hoeffding, Probability inequalities for sums of bounded random variables, Journal of the American Statistical Association 58 (301) (1963) 13–30.

[48] R. Vershynin, Introduction to the Non-Asymptotic Analysis of Random Matrices, in: Y. C. Eldar, G. Kutyniok (Eds.), Compressed Sensing: Theory and Practice, Cambridge University Press, 210–268, 2012.

[49] D. Hsu, S. Kakade, T. Zhang, A tail inequality for quadratic forms of subgaussian random vectors, Electronic Communications in Probability 17 (52) (2012) 1–6.

[50] A. Ehrenfeucht, D. Haussler, M. Kearns, L. Valiant, A general lower bound on the number of examples needed for learning, Information and Computation 82 (3) (1989) 247–261.

[51] M. F. Balcan, P. Long, Active and passive learning of linear separators under log-concave distributions, in: Conference on Learning Theory (COLT), 288–316, 2013.

[52] D. Haussler, N. Littlestone, M. K. Warmuth, Predicting $\{0, 1\}$-functions on randomly drawn points, Information and Computation 115 (2) (1994) 248–292.

[53] R. Vershynin, High-dimensional probability: An introduction with applications in data science, Cambridge University Press, 2018.

Biographies

Berkan Kadıoğluis a PhD candidate at the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, since 2017. He received his B.Sc. (2017) in Electrical and Electronics Engineering from Bilkent University, Ankara, Turkey. His research interests include machine learning, statistical learning, deep learning with a focus on learning from comparisons.

Dr. Peng Tianis a postdoctoral research associate at Northeastern University, Boston, MA. He received his Ph.D. and M.S. degree of Electrical Engineering in 2020 and 2017 respectively, from Northeastern University, Boston, MA. He obtained his B.E. degree of Optoelectronic Information Engineering in 2015 from Huazhong University of Science & and Technology, Wuhan, China. His research interests span learning from comparisons, deep learning and AI for healthcare.

Dr. Jennifer G. Dy is a professor at the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, since 2002. She obtained her MS and PhD in 1997 and 2001 respectively from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, and her BS degree in 1993 from the Department of Electrical Engineering, University of the Philippines. She received an NSF Career award in 2004. She is an editorial board member for the journal, Machine Learning since 2004, publications chair for the International Conference on Machine Learning in 2004, and program committee member for ICML, ACM SIGKDD, AAAI, and SIAM SDM. Her research interests include Machine Learning, Data Mining, Statistical Pattern Recognition, and Computer Vision.

Dr. Deniz Erdoğmus graduated with B.S. in Electrical & Electronics Engineering (EEE), and the B.S. in Mathematics in 1997, and M.S. in EEE in 1999 from the Middle East Technical University, Ankara, Turkey. He received his Ph.D. in Electrical & Computer Engineering from the University of Florida in 2002, where he stayed as a post-doctoral research associate until 2004. Prior to joining the Northeastern faculty in 2008, he held an Assistant Professor position at the Oregon Health and Science University. His expertise is in information theoretic and nonparametric machine learning and adaptive signal processing, specifically focusing on cognitive signal processing including brain interfaces and assistive technologies. Deniz has been serving as an associate editor IEEE Transactions on Signal Processing, Transactions on Neural Networks, Signal Processing Letters, and Elsevier Neurocomputing. He is a member of the IEEE-SPS Machine Learning for Signal Processing Technical Committee.

Dr. Stratis Ioannidis is an associate professor in the Electrical and Computer Engineering Department of Northeastern University, in Boston, MA, where he also holds a courtesy appointment with the College of Computer and Information Science. He received his B.Sc. (2002) in Electrical and Computer Engineering from the National Technical University of Athens, Greece, and his M.Sc. (2004) and Ph.D. (2009) in Computer Science from the University of Toronto, Canada. Prior to joining Northeastern, he was a research scientist at the Technicolor research centers in Paris, France, and Palo Alto, CA, as well as at Yahoo Labs in Sunnyvale, CA. He is the recipient of an NSF CAREER Award, a Google Faculty Research Award, a Facebook Research Award, a Martin W. Essigmann Outstanding Teaching Award, and several best paper awards. His research interests span machine learning, distributed systems, and optimization.

Appendix A. Proof of Lemma 7

We first show an intermediate result. For a random variable $X \sim \mathcal{N}(0, \sigma^2)$ where $\sigma > 0$, we have

$$\begin{aligned}
\mathbb{E}[f(x)] &= \int_{-\infty}^{\infty} f(x) f_X(x) dx = \int_0^{\infty} f(x) f_X(x) dx + \int_0^{\infty} f(-x) f_X(-x) dx \\
&= \int_0^{\infty} f(x) f_X(x) dx + \int_0^{\infty} (1 - f(x)) f_X(x) dx \quad \text{as } f_X \text{ is symmetric around 0,} \\
&= 1/2 + \int_0^{\infty} f(x) f_X(x) dx - \int_0^{\infty} f(x) f_X(x) dx = 1/2. \qquad \text{(A.1)}
\end{aligned}$$

Hence, the expected value of our estimator is

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^{M} Y_m \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right] \quad \text{by Eq.(3),} \\
&= \mathbb{E}\left[Y_m \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right] \qquad \text{(A.2)} \\
&= \mathbb{E}\left[\hat{\boldsymbol{\Sigma}}^{-1}\right] \mathbb{E}\left[Y_m (\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right] \quad \text{by } \hat{\boldsymbol{\Sigma}}^{-1} \perp\!\!\!\perp Y_m, \boldsymbol{X}_{I_m}, \boldsymbol{X}_{J_m}, \\
&= \boldsymbol{\Sigma}^{-1} \mathbb{E}\left[Y_m (\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right] \quad \text{by [41],}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{\Sigma}^{-1}\mathbb{E}\left[(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\mathbb{E}\left[Y_m|\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m}\right]\right] \\
&= \mathbf{\Sigma}^{-1}\mathbb{E}\left[(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\left(2f\left(\boldsymbol{\beta}^\top(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right) - 1\right)\right] \\
&= \mathbf{\Sigma}^{-1}\text{Cov}\left[\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m}, 2f\left(\boldsymbol{\beta}^T(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right) - 1\right] \quad \text{by Eq. (A.1),} \\
&= 2\mathbf{\Sigma}^{-1}\text{Cov}\left[\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m}, \boldsymbol{\beta}^T(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right]\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right)\right] \\
&= 2\mathbf{\Sigma}^{-1}\mathbb{E}\left[(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})^T\right]\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right)\right]\boldsymbol{\beta} \\
&= 4\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right)\right]\boldsymbol{\beta} = 4\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right)\right]\boldsymbol{\beta} \\
&= c_1\boldsymbol{\beta}, \qquad \text{(A.3)}
\end{aligned}
$$

where the fourth to last line is by Lemma 1 and $c_1 = 4\mathbb{E}\left[f'\left(\boldsymbol{\beta}^T(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})\right)\right]$. Note that c_1 is strictly positive as $f(x)$ is non-decreasing, and has limits $\lim_{x\to\infty} f(x) = 1$, $\lim_{x\to-\infty} f(x) = 0$ by Eq. (2). Therefore, there exists an $s \in \mathbb{R}$ at which $f'(s) > 0$. By continuity, $f'(x)$ is therefore strictly positive in an interval around s. As a result, the integral in the expectation which defines c_1 is strictly positive. □

Appendix B. Proof of Lemma 8

Note that, due to Equations (A.2) and (A.3) in Appendix A, we have $c_1\boldsymbol{\beta} = \mathbb{E}[Y_m\mathbf{\Sigma}^{-1}(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})]$. We thus have

$$
\begin{aligned}
\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| &= \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m}) - \mathbb{E}[Y_m\mathbf{\Sigma}^{-1}(\boldsymbol{X}_{I_m} - \boldsymbol{X}_{J_m})]\right\| \\
&= \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu}) - \frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{X}_{J_m} - \boldsymbol{\mu})\right. \\
&\quad \left. - \mathbb{E}[Y_m\mathbf{\Sigma}^{-1}(X_{I_m} - \boldsymbol{\mu})] + \mathbb{E}[Y_m\mathbf{\Sigma}^{-1}(\boldsymbol{X}_{J_m} - \boldsymbol{\mu})]\right\| \\
&\leq \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu}) - \mathbb{E}[Y_m\mathbf{\Sigma}^{-1}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\| \\
&\quad + \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{X}_{J_m} - \boldsymbol{\mu}) - \mathbb{E}[Y_m\mathbf{\Sigma}^{-1}(\boldsymbol{X}_{J_m} - \boldsymbol{\mu})]\right\|,
\end{aligned}
$$

where the last line is by the triangle inequality and the first line is by Eq. (3) and Lemma 7. Then, we show that these terms are bounded by the same probability. We start by defining $Y'_m = -Y_m$ and note that

$$
\begin{aligned}
\Pr(Y'_m \mid \boldsymbol{X}_{I_m} = x, \boldsymbol{X}_{J_m} = y) &= \Pr(-Y_m \mid \boldsymbol{X}_{I_m} = x, \boldsymbol{X}_{J_m} = y) \\
&= 1 - f(\boldsymbol{\beta}^\top(x - y)) \quad \text{by Eq. (1),} \\
&= f(\boldsymbol{\beta}^\top(y - x)) \quad \text{by Eq. (2).} \qquad \text{(B.1)}
\end{aligned}
$$

Then, we have

$$
\Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{X}_{J_m} - \boldsymbol{\mu}) - \mathbb{E}[Y_m\mathbf{\Sigma}^{-1}(\boldsymbol{X}_{J_m} - \boldsymbol{\mu})]\right\| > \epsilon\right)
$$

$$= \Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} -Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{J_m}-\boldsymbol{\mu}) - \mathbb{E}[-Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{J_m}-\boldsymbol{\mu})]\right\| > \epsilon\right)$$

$$= \Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y'_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{J_m}-\boldsymbol{\mu}) - \mathbb{E}[Y'_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{J_m}-\boldsymbol{\mu})]\right\| > \epsilon\right)$$

$$= \Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\| > \epsilon\right), \quad \text{(B.2)}$$

where the last line follows by Eq. (B.1). We use this result to show that

$$\Pr\left(\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| > \epsilon\right) \le \Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|\right.$$

$$\left. + \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{J_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{J_m}-\boldsymbol{\mu})]\right\| > \epsilon\right)$$

$$\le \Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\| > \epsilon/2\right)$$

$$+ \Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{J_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{J_m}-\boldsymbol{\mu})]\right\| > \epsilon/2\right)$$

$$= 2\Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\| > \epsilon/2\right), \quad \text{(B.3)}$$

where the last line follows by Eq. (B.2). Then,

$$\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|$$

$$\le \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu}) - \frac{1}{M}\sum_{m=1}^{M} Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})\right\|$$

$$+ \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|$$

$$\le \left\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\right\| \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})\right\|$$

$$+ \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|$$

$$\le \left\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\right\| \left\|\frac{1}{M}\sum_{m=1}^{M} Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu}) - \mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|$$

$$+ \left\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\right\| \left\|\mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|$$

$$+\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})-\mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|, \tag{B.4}$$

where the first and last inequalities follow by the triangle inequality and the second inequality follows by the Cauchy-Schwarz inequality. Note that,

$$\begin{aligned}
&\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1}\right\|\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})-\mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|>\epsilon\right)\\
&\leq \Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1}\right\|>\sqrt{\epsilon}\right)\\
&+\Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})-\mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|>\sqrt{\epsilon}\right)\\
&\leq \Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1}\right\|>\epsilon\right)\\
&+\Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})-\mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|>\epsilon\right),
\end{aligned}$$

since $\sqrt{\epsilon}>\epsilon$ for $\epsilon<1$. These terms appear in the bound more than once, therefore we can ignore the higher order term by multiplying the lower order terms with a constant number, e.g., 2. This yields

$$\begin{aligned}
&\Pr\left(\left\|\hat{\boldsymbol{\beta}}-c_1\boldsymbol{\beta}\right\|>\epsilon\right)\\
&\leq 2\Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})-\mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|>\epsilon/2\right)\\
&\leq 2\Pr\Bigg(\left\|\hat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1}\right\|\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})-\mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|\\
&+\left\|\hat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1}\right\|\|\mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\|\\
&+\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})-\mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|>\epsilon/2\Bigg)\\
&\leq 2\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1}\right\|\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})-\mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|>\epsilon/6\right)\\
&+2\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1}\right\|\|\mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\|>\epsilon/6\right)\\
&+2\Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})-\mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\right\|>\epsilon/6\right)\\
&\leq 4\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1}\right\|\|\mathbb{E}[Y_m(\boldsymbol{X}_{I_m}-\boldsymbol{\mu})]\|>\epsilon/6\right)
\end{aligned}$$

$$+ 4\Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\| > \epsilon/6\right)$$

$$\leq 4\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\right\| \|\mathbb{E}[Y_m(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\| > \epsilon/6\right)$$

$$+ 4\Pr\left(\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\| > \sqrt{\lambda_d}\epsilon/6\right),$$

for $\epsilon < 1$ where in the last line we divide and multiply by $\boldsymbol{\Sigma}^{-1/2}$ (recall that λ_d is the minimum eigenvalue of $\boldsymbol{\Sigma}$). □

Appendix C. Proof of Lemma 9

We remind the reader that $\boldsymbol{W}_n = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_n - \boldsymbol{\mu})$ and we show that

$$\left\|\frac{1}{M}\sum_{m=1}^{M} Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\|$$

$$= \left\|\frac{1}{M}\sum_{n=1}^{N} \boldsymbol{W}_n \sum_{m: I_m = n} Y_m - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\|$$

$$= \left\|\frac{1}{M}\sum_{n=1}^{N} \boldsymbol{W}_n M_n \mathbb{E}[Y_m | I_m = n, \{\boldsymbol{X}_{n'} = \boldsymbol{x}_{n'}\}_{n'=1}^{N}] - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right.$$

$$\left. + \frac{1}{M}\sum_{n=1}^{N} \boldsymbol{W}_n M_n \left[\frac{1}{M_n}\sum_{m: I_m = n} Y_m - \mathbb{E}[Y_m | I_m = n, \{\boldsymbol{X}_{n'} = \boldsymbol{x}_{n'}\}_{n'=1}^{N}]\right]\right\|$$

$$\leq \left\|\frac{1}{M}\sum_{n=1}^{N} \boldsymbol{W}_n M_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\| \quad \text{by Eq. (9),}$$

$$+ \left\|\frac{1}{M}\sum_{n=1}^{N} \boldsymbol{W}_n M_n \Delta_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N})\right\|, \tag{C.1}$$

where the last line follows by Eq. (11) and the triangle inequality. We expand the first term in Eq. (C.1) as follows:

$$\left\|\frac{1}{M}\sum_{n=1}^{N} \boldsymbol{W}_n M_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\|$$

$$= \left\|\frac{1}{M}\sum_{n=1}^{N} \boldsymbol{W}_n M_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) - \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N})\right.$$

$$\left. + \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\|$$

$$
\begin{aligned}
&\leq \left\| \frac{1}{M} \sum_{n=1}^{N} \boldsymbol{W}_n M_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) - \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| \\
&+ \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})] \right\| \\
&= \left\| \sum_{n=1}^{N} \left(\frac{M_n}{M} - \frac{1}{N} \right) \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| \\
&+ \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})] \right\| . \qquad \text{(C.2)}
\end{aligned}
$$

For the second term in Eq. (C.2), we have

$$
\begin{aligned}
&\left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})] \right\| \\
&\leq \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n \big(\tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) - g_n(\boldsymbol{X}_n) \big) \right\| \\
&+ \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n g_n(\boldsymbol{X}_n) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})] \right\| \\
&= \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n z_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| \\
&+ \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n g_n(\boldsymbol{X}_n) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})] \right\| , \qquad \text{(C.3)}
\end{aligned}
$$

by Eq. (10). By combining Equations (C.1), (C.2), (C.3) we get

$$
\begin{aligned}
&\left\| \frac{1}{M} \sum_{m=1}^{M} Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu}) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})] \right\| \\
&\leq \left\| \frac{1}{M} \sum_{n=1}^{N} \boldsymbol{W}_n M_n \Delta_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| + \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n z_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| \\
&+ \left\| \sum_{n=1}^{N} \left(\frac{M_n}{M} - \frac{1}{N} \right) \boldsymbol{W}_n \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}) \right\| \\
&+ \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{W}_n g_n(\boldsymbol{X}_n) - \mathbb{E}[Y_m \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})] \right\| . \quad \square
\end{aligned}
$$

Appendix D. Proof of Lemma 10

The term of interest is

$$
\begin{aligned}
&\Pr\left(\left\|\sum_{n=1}^{N}\left(\frac{M_n}{M}-\frac{1}{N}\right)\boldsymbol{W}_n\tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N})\right\| > \epsilon\right)\\
&\leq \Pr\left(\sum_{n=1}^{N}\left\|\left(\frac{M_n}{M}-\frac{1}{N}\right)\boldsymbol{W}_n\tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N})\right\| \geq \epsilon\right) \quad \text{by the triangle inequality,}\\
&\leq \Pr\left(\sum_{n=1}^{N}\left|\frac{M_n}{M}-\frac{1}{N}\right|\|\boldsymbol{W}_n\| \geq \epsilon\right) \quad \text{by the fact that } \left|\tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^{N})\right| \leq 1,\\
&= \Pr\left(\sum_{n=1}^{N}\left|\frac{M_n}{M}-\frac{1}{N}\right|\|\boldsymbol{W}_n\| > \epsilon \mid \cap_{n=1}^{N}\left\{\|\boldsymbol{W}_n\| \leq \sqrt{\delta_0}\right\}\right)\times\\
&\Pr\left(\cap_{n=1}^{N}\left\{\|\boldsymbol{W}_n\| \leq \sqrt{\delta_0}\right\}\right)\\
&+\Pr\left(\sum_{n=1}^{N}\left|\frac{M_n}{M}-\frac{1}{N}\right|\|\boldsymbol{W}_n\| > \epsilon \mid \cup_{n=1}^{N}\left\{\|\boldsymbol{W}_n\| \geq \sqrt{\delta_0}\right\}\right)\times\\
&\Pr\left(\cup_{n=1}^{N}\left\{\|\boldsymbol{W}_n\| \geq \sqrt{\delta_0}\right\}\right)\\
&\leq \Pr\left(\left\{\sum_{n=1}^{N}\left|\frac{M_n}{M}-\frac{1}{N}\right|\|\boldsymbol{W}_n\| > \epsilon\right\}\cap\cap_{n=1}^{N}\left\{\|\boldsymbol{W}_n\| \leq \sqrt{\delta_0}\right\}\right)\\
&+\Pr\left(\cup_{n=1}^{N}\left\{\|\boldsymbol{W}_n\| \geq \sqrt{\delta_0}\right\}\right)\\
&\leq \Pr\left(\sum_{n=1}^{N}\left|\frac{M_n}{M}-\frac{1}{N}\right| > \epsilon/\sqrt{\delta_0}\right)+\sum_{n=1}^{N}\Pr\left(\|\boldsymbol{W}_n\| \geq \sqrt{\delta_0}\right), \qquad \text{(D.1)}
\end{aligned}
$$

where the last line is by a union bound and letting $A = \sum_{n=1}^{N}\left|\frac{M_n}{M}-\frac{1}{N}\right|\|\boldsymbol{W}_n\| > \epsilon \cap \cap_{n=1}^{N}\left\{\|\boldsymbol{W}_n\| \leq \sqrt{\delta_0}\right\}$ and noticing that the event A implies the event $B = \sum_{n=1}^{N}\left|\frac{M_n}{M}-\frac{1}{N}\right| > \epsilon/\sqrt{\delta_0}$ and therefore $A \subseteq B$ and $\Pr(A) \leq \Pr(B)$. Since M_n are binomial distributed with parameter $1/N$, we have

$$
\Pr\left(\sum_{n=1}^{N}\left|\frac{M_n}{M}-\frac{1}{N}\right| > \epsilon/\sqrt{\delta_0}\right) \leq 2^N e^{-\frac{\epsilon^2 M}{2\delta_0}} \quad \text{by Lemma 3.} \qquad \text{(D.2)}
$$

As $\|\boldsymbol{W}_n\|^2$ is centralized chi-squared distributed with d degrees of freedom,

$$
\Pr\left(\|\boldsymbol{W}_n\| > \sqrt{\delta_0}\right) = \Pr\left(\|\boldsymbol{W}_n\|^2 > \delta_0\right) \leq \left(\frac{\delta_0}{d}e^{1-\frac{\delta_0}{d}}\right)^{d/2}, \qquad \text{(D.3)}
$$

for $\delta_0 > d$ by Lemma 2. By combining Equations (D.1), (D.2) and (D.3), we get

$$\Pr\left(\left\|\sum_{n=1}^{N}\left(\frac{M_n}{M}-\frac{1}{N}\right)\boldsymbol{W}_n\tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^N)\right\| > \epsilon\right)$$
$$\leq 2^N e^{-\frac{\epsilon^2 M}{2\delta_0}} + N\Big(\frac{\delta_0}{d}e^{1-\frac{\delta_0}{d}}\Big)^{d/2}. \quad \square \tag{D.4}$$

Appendix E. Proof of Lemma 11

We omit the dependence on $\{\boldsymbol{X}_{n'}\}_{n'=1}^N$ for $\Delta_n\left(\{\boldsymbol{X}_{n'}\}_{n'=1}^N\right)$ and note that

$$\Pr\left(\left\|\frac{1}{M}\sum_{n=1}^{N}\boldsymbol{W}_n M_n\Delta_n\right\| > \epsilon\right) \leq \Pr\left(\sum_{n=1}^{N}\left\|\frac{1}{M}\boldsymbol{W}_n M_n\Delta_n\right\| > \epsilon\right)$$
$$= \Pr\left(\sum_{n=1}^{N}\left\|\frac{1}{M}\boldsymbol{W}_n M_n\right\| |\Delta_n| > \epsilon\right)$$
$$= \Pr\left(\sum_{n=1}^{N}\left\|\frac{1}{M}\boldsymbol{W}_n M_n\right\| |\Delta_n| > \epsilon \mid \cap_{n=1}^N \left\{|\Delta_n| < \sqrt{\delta_1}\right\}\right)\times$$
$$\Pr\left(\cap_{n=1}^N \left\{|\Delta_n| < \sqrt{\delta_1}\right\}\right)$$
$$+ \Pr\left(\sum_{n=1}^{N}\left\|\frac{1}{M}\boldsymbol{W}_n M_n\right\| |\Delta_n| > \epsilon \mid \cup_{n=1}^N \left\{|\Delta_n| \geq \sqrt{\delta_1}\right\}\right)\times$$
$$\Pr\left(\cup_{n=1}^N \left\{|\Delta_n| \geq \sqrt{\delta_1}\right\}\right)$$
$$\leq \Pr\left(\left\{\sum_{n=1}^{N}\left\|\frac{1}{M}\boldsymbol{W}_n M_n\right\| |\Delta_n| > \epsilon\right\} \cap \cap_{n=1}^N \left\{|\Delta_n| < \sqrt{\delta_1}\right\}\right)$$
$$+ \Pr\left(\cup_{n=1}^N \left\{|\Delta_n| \geq \sqrt{\delta_1}\right\}\right)$$
$$\leq \Pr\left(\sum_{n=1}^{N}\left\|\frac{1}{M}\boldsymbol{W}_n M_n\right\| > \epsilon/\sqrt{\delta_1}\right) + N\Pr\left(\left\{|\Delta_n| \geq \sqrt{\delta_1}\right\}\right), \tag{E.1}$$

where we follow a similar approach as in Eq. (D.1). The first term in Eq. (E.1) can be expanded as

$$\Pr\Big(\sum_{n=1}^{N}\Big\|\frac{1}{M}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_n-\boldsymbol{\mu})M_n\Big\| > \frac{\epsilon}{\sqrt{\delta_1}}\Big)$$
$$= \Pr\left(\sum_{n=1}^{N}\left\|\frac{M_n}{M}\boldsymbol{W}_n - \frac{1}{N}\boldsymbol{W}_n + \frac{1}{N}\boldsymbol{W}_n\right\| \geq \frac{\epsilon}{\sqrt{\delta_1}}\right)$$

$$\leq \Pr\left(\sum_{n=1}^{N}\left\|\left(\frac{M_n}{M}-\frac{1}{N}\right)\boldsymbol{W}_n\right\| + \sum_{n=1}^{N}\left\|\frac{1}{N}\boldsymbol{W}_n\right\| \geq \frac{\epsilon}{\sqrt{\delta_1}}\right)$$
$$\leq \Pr\left(\sum_{n=1}^{N}\left\|\left(\frac{M_n}{M}-\frac{1}{N}\right)\boldsymbol{W}_n\right\| \geq \frac{\epsilon}{2\sqrt{\delta_1}}\right) + \Pr\left(\sum_{n=1}^{N}\left\|\frac{1}{N}\boldsymbol{W}_n\right\| \geq \frac{\epsilon}{2\sqrt{\delta_1}}\right). \tag{E.2}$$

We have

$$\Pr\left(\sum_{n=1}^{N}\left\|\left(\frac{M_n}{M}-\frac{1}{N}\right)\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_n-\boldsymbol{\mu})\right\| \geq \frac{\epsilon}{2\sqrt{\delta_1}}\right) \leq 2^N e^{-\frac{\epsilon^2 M}{8\delta_1\delta_2}} + N\left(\frac{\delta_2}{d}e^{1-\frac{\delta_2}{d}}\right)^{d/2} \tag{E.3}$$

by Eq. (D.4). Then,

$$\Pr\left(\sum_{n=1}^{N}\left\|\frac{1}{N}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_n-\boldsymbol{\mu})\right\| \geq \frac{\epsilon}{2\sqrt{\delta_1}}\right) \leq \sum_{n=1}^{N}\Pr\left(\left\|\frac{1}{N}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_n-\boldsymbol{\mu})\right\| \geq \frac{\epsilon}{2N\sqrt{\delta_1}}\right)$$
$$= \sum_{n=1}^{N}\Pr\left(\|\boldsymbol{W}_n\| \geq \frac{\epsilon}{2\sqrt{\delta_1}}\right) \leq \sum_{n=1}^{N}\left(\frac{\epsilon^2}{4d\delta_1}e^{1-\frac{\epsilon^2}{4d\delta_1}}\right)^{d/2} \text{ by Lemma 2,}$$
$$= N\left(\frac{\epsilon^2}{4d\delta_1}e^{1-\frac{\epsilon^2}{4d\delta_1}}\right)^{d/2}, \tag{E.4}$$

where the first line holds by a union bound. The second term in Eq. (E.1) is bounded by

$$\Pr\left(\left|\Delta_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^N)\right| \geq \sqrt{\delta_1}\right)$$
$$= \Pr\left(\left|\frac{1}{M_n}\sum_{m:I_m=n} Y_m - \tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^N)\right| \geq \sqrt{\delta_1}\right) \quad \text{by Eq. (11),}$$
$$= \sum_{k=0}^{M}\left[\int \Pr\left(\left|\frac{1}{M_n}\sum_{m:I_m=n} Y_m - \tilde{g}_n(\{\boldsymbol{x}_{n'}\}_{n'=1}^N)\right| \geq \sqrt{\delta_1} \mid M_n = k, \{\boldsymbol{X}_{n'} = \boldsymbol{x}_{n'}\}_{n'=1}^N\right)\right.$$
$$\left.\prod_{n'=1}^{N} f_{\boldsymbol{X}}(\boldsymbol{x}_{n'})d\boldsymbol{x}_{n'}\right] \cdot \Pr(M_n = k). \tag{E.5}$$

Due to conditioning on $\{\boldsymbol{X}_{n'} = \boldsymbol{x}_{n'}\}_{n'=1}^N$, labels Y_m are independent. Therefore,

$$\Pr\left(\left|\frac{1}{M_n}\sum_{m:I_m=n} Y_m - \tilde{g}_n(\{\boldsymbol{x}_{n'}\}_{n'=1}^N)\right| \geq \sqrt{\delta_1} \mid M_n = k, \{\boldsymbol{X}_{n'} = \boldsymbol{x}_{n'}\}_{n'=1}^N\right)$$
$$= \Pr\left(\left|\frac{1}{k}\sum_{m:I_m=n} Y_m - \tilde{g}_n(\{\boldsymbol{x}_{n'}\}_{n'=1}^N)\right| \geq \sqrt{\delta_1} \mid \{\boldsymbol{X}_{n'} = \boldsymbol{x}_{n'}\}_{n'=1}^N\right) \leq 2e^{-\frac{k\delta_1}{2}},$$

where the last line is by Lemma 4. Substituting this result back into Eq. (E.5) yields

$$\Pr\left(\left|\Delta_n\right| \geq \sqrt{\delta}_1\right) \leq \sum_{k=0}^{M} \int 2e^{-\frac{k\delta_1}{2}} \Big(\prod_{n'=1}^{N} f_{\boldsymbol{X}}(\boldsymbol{x}_{n'}) d\boldsymbol{x}_{n'}\Big) \Pr(M_n = k)$$
$$= 2\sum_{k=0}^{M} e^{-\frac{k\delta_1}{2}} \Pr(M_n = k). \tag{E.6}$$

By construction, $M_n, n \in [N]$ are binomial distributed with number of trials M and $p = \frac{1}{N}$. The moment generating function of M_n is $(1 - p + pe^t)^M$. The summation in Eq. (E.6) is the moment generating function of M_n with $t = -\delta_1/2$, which yields

$$\sum_{n=1}^{N} \Pr\left(\left|\Delta_n\right| \geq \sqrt{\delta_1}\right) \leq 2N\left(1 - \frac{1}{N} + \frac{1}{N}e^{-\frac{\delta_1}{2}}\right)^M. \tag{E.7}$$

We want to use an equivalent of this quantity in the form of an exponential since it will be easier to compare it with other terms. For this reason, we use the following lemma.

Lemma 15. *Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be positive sequences such that $a_n \to \infty$. Let $\alpha \in \mathbb{R}$. Then,*

$$\left(1 + \frac{\alpha}{a_n}\right)^{b_n} = e^{\frac{\alpha b_n}{a_n} + o\left(\frac{\alpha b_n}{a_n}\right)}. \tag{E.8}$$

Proof. By the Taylor approximation $\log(1 + x) = x + o(x)$ as $x \to 0$,

$$\left(1 + \frac{\alpha}{a_n}\right)^{b_n} = e^{b_n \log\left(1 + \frac{\alpha}{a_n}\right)} = e^{b_n\left(\frac{\alpha}{a_n} + o\left(\frac{\alpha}{a_n}\right)\right)} = e^{\frac{\alpha b_n}{a_n} + o\left(\frac{\alpha b_n}{a_n}\right)}.$$

□

Recall the Taylor expansion $e^x = 1 + x + o(x)$ as $x \to 0$ and note that

$$2N\left(1 - \frac{1}{N} + \frac{1}{N}e^{-\frac{\delta_1}{2}}\right)^M = 2N\left(1 + \frac{1}{N}\left(e^{-\frac{\delta_1}{2}} - 1\right)\right)^M$$
$$= 2N\left(1 + \frac{1}{N}\left(-\frac{\delta_1}{2} - o\left(\frac{\delta_1}{2}\right)\right)\right)^M + 2N\left(1 + \left(-\frac{\delta_1}{2N} - o\left(\frac{\delta_1}{2N}\right)\right)\right)^M$$
$$= 2Ne^{-\frac{M\delta_1}{2N} - o\left(\frac{M\delta_1}{2N}\right)} = 2e^{\log N - \frac{M\delta_1}{2N} - o\left(\frac{M\delta_1}{2N}\right)}, \tag{E.9}$$

where the second line is by the Taylor expansion of e^x and the last line is by Lemma 15. Combining Equations (E.1), (E.2), (E.3), (E.4), (E.7), (E.9), we get

$$\Pr\left(\left\|\frac{1}{M}\sum_{n=1}^{N} \boldsymbol{W}_n M_n \Delta_n\right\| > \epsilon\right) \leq 2^N e^{-\frac{\epsilon^2 M}{8\delta_1\delta_2}} + N\left(\frac{\delta_2}{d}e^{1-\frac{\delta_2}{d}}\right)^{d/2}$$

$$+ N\left(\frac{\epsilon^2}{4d\delta_1}e^{1-\frac{\epsilon^2}{4d\delta_1}}\right)^{d/2} + 2e^{\log N - \frac{M\delta_1}{2N} - o\left(\frac{M\delta_1}{2N}\right)}. \quad \square$$

Appendix F. Proof of Lemma 12

For brevity, we use z_n instead of $z_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^N)$ below. Note that

$$\begin{aligned}
&\Pr\left(\left\|\frac{1}{N}\sum_{n=1}^N \boldsymbol{W}_n z_n\right\| > \epsilon\right) \le \Pr\left(\sum_{n=1}^N \left\|\frac{1}{N}\boldsymbol{W}_n z_n\right\| > \epsilon\right) \\
&= \Pr\left(\sum_{n=1}^N \left\|\frac{1}{N}\boldsymbol{W}_n\right\| |z_n| > \epsilon \mid \cap_{n=1}^N \left\{|z_n| < \sqrt{\delta_3}\right\}\right)\times \\
&\Pr\left(\cap_{n=1}^N \left\{|z_n| < \sqrt{\delta_3}\right\}\right) \\
&+ \Pr\left(\sum_{n=1}^N \left\|\frac{1}{N}\boldsymbol{W}_n\right\| |z_n| > \epsilon \mid \cup_{n=1}^N \left\{|z_n| \ge \sqrt{\delta_3}\right\}\right)\times \\
&\Pr\left(\cup_{n=1}^N \left\{|z_n| \ge \sqrt{\delta_3}\right\}\right) \\
&\le \Pr\left(\left\{\sum_{n=1}^N \left\|\frac{1}{N}\boldsymbol{W}_n\right\| |z_n| > \epsilon\right\} \cap \cap_{n=1}^N \left\{|z_n| < \sqrt{\delta_3}\right\}\right) \\
&+ \Pr\left(\cup_{n=1}^N \left\{|z_n| \ge \sqrt{\delta_3}\right\}\right) \\
&\le \Pr\left(\sum_{n=1}^N \left\|\frac{1}{N}\boldsymbol{W}_n\right\| > \epsilon/\sqrt{\delta_3}\right) + N\Pr\left(|z_n| \ge \sqrt{\delta_3}\right), \qquad \text{(F.1)}
\end{aligned}$$

where the first line follows by the triangle inequality and by following a similar approach as in Eq. (D.1). We have

$$\Pr\left(\sum_{n=1}^N \left\|\frac{1}{N}\boldsymbol{W}_n\right\| > \epsilon/\sqrt{\delta_3}\right) \le N\left(\frac{\epsilon^2}{d\delta_3}e^{1-\frac{\epsilon^2}{d\delta_3}}\right)^{d/2}, \qquad \text{(F.2)}$$

by Lemma 2 for $\delta_3 \le \epsilon^2/d$. We also have

$$\begin{aligned}
&\Pr\left(\left|z_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^N)\right| \ge \sqrt{\delta_3}\right) \\
&= \int \Pr\left(\left|z_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^N)\right| \ge \sqrt{\delta_3} \mid \boldsymbol{X}_n = \boldsymbol{x}\right) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}. \qquad \text{(F.3)}
\end{aligned}$$

For the probability inside the integral, we have

$$\Pr\left(\left|z_n\{\boldsymbol{X}_{n'}\}_{n'=1}^N\right| \ge \sqrt{\delta_3} \mid \boldsymbol{X}_n = \boldsymbol{x}\right)$$

$$
\begin{aligned}
&= \Pr\left(\left|\tilde{g}_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^N) - g_n(\boldsymbol{X}_n)\right| \geq \sqrt{\delta_3} \mid \boldsymbol{X}_n = \boldsymbol{x}\right) \\
&= \Pr\left(\left|\sum_{j=1}^N g_{n,j}(\boldsymbol{x}, \boldsymbol{X}_j) \Pr(J=j) - g_n(\boldsymbol{x})\right| \geq \sqrt{\delta_3}\right) \\
&= \Pr\left(\left|\frac{1}{N}\sum_{j=1}^N g_{n,j}(\boldsymbol{x}, \boldsymbol{X}_j) - g_n(\boldsymbol{x})\right| \geq \sqrt{\delta_3}\right),
\end{aligned}
\tag{F.4}
$$

where the first line follows by Eq. (10). Notice that

$$
\begin{aligned}
g_n(\boldsymbol{x}) &= \mathbb{E}[Y_m \mid I_M = n, \boldsymbol{X}_n = \boldsymbol{x}] \\
&= \sum_{j=1}^N \mathbb{E}[Y_m \mid I_m = n, J_m = j, \boldsymbol{X}_n = \boldsymbol{x}] \Pr(J_m = j) \\
&= \frac{1}{N}\sum_{j=1}^N \int \mathbb{E}[Y_m \mid I_m = n, J_m = j, \boldsymbol{X}_n = \boldsymbol{x}, \boldsymbol{X}_j = y] f_{\boldsymbol{X}_j}(\boldsymbol{y}) d\boldsymbol{y} \\
&= \frac{1}{N} g_{n,n}(\boldsymbol{x}, \boldsymbol{x}) + \frac{1}{N}\sum_{j\in[N]\setminus n} \int g_{n,j}(\boldsymbol{x}, \boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{y}) d\boldsymbol{y} \\
&= \frac{1}{N} g_{n,n}(\boldsymbol{x}, \boldsymbol{x}) + \frac{N-1}{N} \int g_{n,j}(\boldsymbol{x}, \boldsymbol{y}) f_{\boldsymbol{X}_j}(\boldsymbol{y}) d\boldsymbol{y}.
\end{aligned}
$$

Then we continue with

$$
\begin{aligned}
&\Pr\left(\left|\frac{1}{N}\sum_{j=1}^N g_{n,j}(\boldsymbol{x}, \boldsymbol{X}_j) - g_n(\boldsymbol{x})\right| \geq \sqrt{\delta_3}\right) \\
&= \Pr\left(\left|\frac{1}{N} g_{n,n}(\boldsymbol{x}, \boldsymbol{x}) + \frac{1}{N}\sum_{j\in[N]\setminus n} g_{n,j}(\boldsymbol{x}, \boldsymbol{X}_j)\right.\right. \\
&\left.\left. - \frac{1}{N} g_{n,n}(\boldsymbol{x}, \boldsymbol{x}) - \frac{1}{N}\sum_{j\in[N]\setminus n} \int g_{n,j}(\boldsymbol{x}, \boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{y}) d\boldsymbol{y}\right| \geq \sqrt{\delta_3}\right) \\
&= \Pr\left(\left|\frac{1}{N}\sum_{j\in[N]\setminus n} g_{n,j}(\boldsymbol{x}, \boldsymbol{X}_j) - \frac{N-1}{N}\int g_{n,j}(\boldsymbol{x}, \boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{y}) d\boldsymbol{y}\right| \geq \sqrt{\delta_3}\right) \\
&= \Pr\left(\left|\frac{1}{N-1}\sum_{j\in[N]\setminus n} g_{n,j}(\boldsymbol{x}, \boldsymbol{X}_j) - \int g_{n,j}(\boldsymbol{x}, \boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{y}) d\boldsymbol{y}\right| \geq \frac{N}{N-1}\sqrt{\delta_3}\right) \\
&\leq 2e^{-\frac{N^3\delta_3}{2(N-1)^2}} \leq 2e^{-\frac{N\delta_3}{2}},
\end{aligned}
\tag{F.5}
$$

where the last line is due to Lemma 4 and the fact that $1 < N/(N-1) \leq 2$ for $N > 1$. Substituting Eq. (F.5) into Eq. (F.3) gives

$$
\Pr\left(\left|z_n(\{\boldsymbol{X}_{n'}\}_{n'=1}^N)\right| \geq \sqrt{\delta_3}\right) \leq \int 2e^{-\frac{N\delta_3}{2}} f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = 2e^{-\frac{N\delta_3}{2}}. \tag{F.6}
$$

By combining Equations (F.1), (F.2) and (F.6), we get

$$\Pr\left(\left\|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{W}_n z_n\{\boldsymbol{X}_{n'}\}_{n'=1}^{N}\right\| > \epsilon\right) \leq N\left(\frac{\epsilon^2}{d\delta_3}e^{1-\frac{\epsilon^2}{d\delta_3}}\right)^{d/2} + 2Ne^{-\frac{N\delta_3}{2}}. \quad \square$$

Appendix G. Proof of Lemma 13

We have,

$$\Pr\left(\left\|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{W}_n g_n(\boldsymbol{X}_n) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\| > \epsilon\right). \tag{G.1}$$

We first prove that $\boldsymbol{W}_n g(\boldsymbol{X}_n)$ is sub-gaussian. By Proposition 2.5.2 (iv) of Vershynin [53], for every sub-gaussian random variable X, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[e^{X^2/C}\right] \leq 2. \tag{G.2}$$

Let $\boldsymbol{v} \in S^{d-1}$ and $W_n = \boldsymbol{v}^\top \boldsymbol{W}_n \sim \mathcal{N}(\boldsymbol{0}, 1)$. Note that W_n is sub-gaussian and $|g_n(\boldsymbol{X}_n)| \leq 1$. Then, for $s > 0$ we have

$$\begin{aligned}
&\Pr\left(|\boldsymbol{v}^\top \boldsymbol{W}_n g(\boldsymbol{X}_n)| > t\right) = \Pr\left(s\left(\boldsymbol{v}^\top \boldsymbol{W}_n g(\boldsymbol{X}_n)\right)^2 > st^2\right)\\
&= \Pr\left(e^{s\left(\boldsymbol{v}^\top \boldsymbol{W}_n g(\boldsymbol{X}_n)\right)^2} > e^{st^2}\right)\\
&\leq e^{-st^2}\mathbb{E}\left[e^{s\left(\boldsymbol{v}^\top \boldsymbol{W}_n g(\boldsymbol{X}_n)\right)^2}\right] \quad \text{by Markov's inequality,}\\
&\leq e^{-st^2}\mathbb{E}\left[e^{s\left(\boldsymbol{v}^\top \boldsymbol{W}_n\right)^2}\right] = e^{-st^2}\mathbb{E}\left[e^{sW_n^2}\right] \leq 2e^{-\frac{t^2}{C}},
\end{aligned} \tag{G.3}$$

where the last line holds by Eq. (G.2) for an appropriate constant $C > 0$ and by setting $s = 1/C$. As the tail of $\boldsymbol{v}^\top \boldsymbol{W}_n g(\boldsymbol{X}_n)$ decays super-exponentially for all $\boldsymbol{v} \in S^{d-1}$, $\boldsymbol{\xi}_n = \boldsymbol{W}_n g_n(\boldsymbol{X}_n) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]$ is indeed sub-gaussian for $n \in [N]$. From Proposition 2.6.1 of [53], we know that the sum of independent zero-mean sub-gaussian random variables is sub-gaussian and

$$\left\|\sum_{n=1}^{N}\boldsymbol{\xi}_n\right\|_{\psi_2}^2 \leq c'\sum_{n=1}^{N}\|\boldsymbol{\xi}_n\|_{\psi_2}^2, \tag{G.4}$$

where $c' > 0$ is a constant. We have

$$\begin{aligned}
&\Pr\left(\left\|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{W}_n g_n(\boldsymbol{X}_n) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right\| > \epsilon\right)\\
&= \Pr\left(\left\|\sum_{n=1}^{N}\left(\boldsymbol{W}_n g_n(\boldsymbol{X}_n) - \mathbb{E}[Y_m\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\right)\right\| > N\epsilon\right)
\end{aligned}$$

$$= \Pr\left(\left\|\sum_{n=1}^{N}\boldsymbol{\xi}_n\right\| > N\epsilon\right) \le e^{-\frac{1}{4}\left(\sqrt{\frac{2N^2\epsilon^2}{\|\sum_{n=1}^{N}\boldsymbol{\xi}_n\|_{\psi_2}^2}-d}-\sqrt{d}\right)^2} \le e^{-\frac{1}{4}\left(\sqrt{\frac{N\epsilon^2}{c'\|\boldsymbol{\xi}_n\|_{\psi_2}^2}-d}-\sqrt{d}\right)^2}$$
$$= e^{-\frac{1}{4}\left(\sqrt{\frac{N\epsilon^2}{c_2}-d}-\sqrt{d}\right)^2},$$

where the last lines follow by Lemma 6, Eq. (G.4) and we denote $c_2 = c' \|\boldsymbol{\xi}_n\|_{\psi_2}^2 > 0$; note that this an absolute constant does not depend on d or N. □

Appendix H. Proof of Lemma 14

We define $\boldsymbol{W} = [\boldsymbol{W}_{N+1}, \ldots, \boldsymbol{W}_{2N}]^\top \in \mathbb{R}^{N\times d}$ such that $\boldsymbol{W}_i = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu})$ where $i \in [2N]/[N]$. Also, let $\lambda_{\min}[\boldsymbol{A}]$ be the minimum singular value of a matrix $\boldsymbol{A}$. We first prove that the minimum singular value of sum of symmetric matrices is lower bounded.

Lemma 16. *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$ be symmetric matrices. Then,*

$$\lambda_{\min}[\boldsymbol{A}] + \lambda_{\min}[\boldsymbol{B}] \le \lambda_{\min}[\boldsymbol{A}+\boldsymbol{B}].$$

Proof. We have,

$$\begin{aligned}\lambda_{\min}[\boldsymbol{A}+\boldsymbol{B}] &= \min_{\boldsymbol{x}\in\mathbb{R}^d}\left\{\frac{\boldsymbol{x}^\top(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}}{\boldsymbol{x}^\top\boldsymbol{x}}\right\} = \min_{\boldsymbol{x}\in\mathbb{R}^d}\left\{\frac{\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top\boldsymbol{x}} + \frac{\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top\boldsymbol{x}}\right\}\\ &\ge \min_{\boldsymbol{x}\in\mathbb{R}^d}\left\{\frac{\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top\boldsymbol{x}}\right\} + \min_{\boldsymbol{y}\in\mathbb{R}^d}\left\{\frac{\boldsymbol{y}^\top\boldsymbol{B}\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}}\right\} = \lambda_{\min}[\boldsymbol{A}] + \lambda_{\min}[\boldsymbol{B}].\end{aligned}$$

□

Hence,

$$\begin{aligned}&\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\right\| > \epsilon\right) \le \Pr\left(\left\|\boldsymbol{\Sigma}^{-1/2}\right\| \cdot \left\|\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{1/2} - \boldsymbol{I}\right\| \cdot \left\|\boldsymbol{\Sigma}^{-1/2}\right\| > \epsilon\right)\\
&= \Pr\left(\lambda_d^{-1}\left\|\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{1/2} - \boldsymbol{I}\right\| > \epsilon\right) = \Pr\left(\left\|\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{1/2}\right\| > \lambda_d\epsilon + 1\right)\\
&= \Pr\left(\left\|\left(\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\right)^{-1}\right\| > \lambda_d\epsilon + 1\right)\\
&= \Pr\left(\lambda_{\min}\left[\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\right] < \frac{1}{\lambda_d\epsilon+1}\right)\\
&= \Pr\left(\lambda_{\min}\left[\sum_{i=1}^{N}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\top\boldsymbol{\Sigma}^{-1/2}\right] < \frac{N-d-2}{\lambda_d\epsilon+1}\right) \quad \text{by Eq. (4),}\\
&= \Pr\Bigg(\lambda_{\min}\Bigg[\sum_{i=1}^{N}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1/2}\\
&+ \sum_{i=1}^{N}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top\boldsymbol{\Sigma}^{-1/2} + \sum_{i=1}^{N}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\top\boldsymbol{\Sigma}^{-1/2}\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1/2} \Bigg] < \frac{N-d-2}{\lambda_d \epsilon + 1} \Bigg) \\
&= \Pr\Bigg(\lambda_{\min} \Bigg[\sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1/2} \\
&+ \sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1/2} \Bigg] < \frac{N-d-2}{\lambda_d \epsilon + 1} \Bigg) \\
&\leq \Pr\Bigg(\lambda_{\min} \Bigg[\sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1/2} \Bigg] \\
&+ \lambda_{\min} \Bigg[\sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1/2} \Bigg] < \frac{N-d-2}{\lambda_d \epsilon + 1} \Bigg) \quad \text{by Lemma 16,} \\
&= \Pr\Bigg(\lambda_{\min} \Bigg[\sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1/2} \Bigg] < \frac{N-d-2}{\lambda_d \epsilon + 1} \Bigg) \\
&= \Pr\Bigg(\lambda_{\min} \Big[\boldsymbol{W}^\top \boldsymbol{W} \Big] < \frac{N-d-2}{\lambda_d \epsilon + 1} \Bigg) = \Pr\Bigg(\lambda_{\min} \Big[\boldsymbol{W} \Big] < \sqrt{\frac{N-d-2}{\lambda_d \epsilon + 1}} \Bigg).
\end{aligned}
$$

Notice that the rows of $\boldsymbol{W}$ are independent isotropic Gaussian vectors. Therefore, we can apply Lemma 5 to get,

$$
\begin{aligned}
\Pr\left(\left\| \hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} \right\| > \epsilon \right) &\leq \Pr\left(\lambda_{\min}\left[\boldsymbol{W}\right] < \sqrt{\frac{N-d-2}{\lambda_d \epsilon + 1}} \right) \\
&\leq 2e^{-c_4\left(\sqrt{N} - \sqrt{\frac{N-d-2}{\lambda_d \epsilon + 1}} - c_3\sqrt{d}\right)^2}, \qquad \text{(H.1)}
\end{aligned}
$$

for $\sqrt{N} > \sqrt{\frac{N-d-2}{\lambda_d \epsilon + 1}} + c_3\sqrt{d}$ where $c_3, c_4 > 0$ are constants. Furthermore, we have

$$
\begin{aligned}
&\|\mathbb{E}[Y_m(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\| \leq \mathbb{E}[\|Y_m(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})\|] \quad \text{by Jensen's inequality,} \\
&\leq \mathbb{E}[\|\boldsymbol{X}_{I_m} - \boldsymbol{\mu}\|] \leq \mathbb{E}\left[\left\| \boldsymbol{\Sigma}^{1/2} \right\| \cdot \left\| \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_{I_m} - \boldsymbol{\mu}) \right\| \right] \\
&\leq \sqrt{\lambda_1} \mathbb{E}[\|\boldsymbol{W}_{I_m}\|] = \sqrt{2\lambda_1} \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&\Pr\left(\left\| \hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} \right\| \cdot \|\mathbb{E}[Y_m(\boldsymbol{X}_{I_m} - \boldsymbol{\mu})]\| > \epsilon \right) \\
&\leq \Pr\left(\left\| \hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} \right\| \sqrt{2\lambda_1} \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} > \epsilon \right) \leq 2e^{-c_4\left(\sqrt{N} - \sqrt{\frac{N-d-2}{\frac{\Gamma\left(\frac{d}{2}\right)\lambda_d}{\Gamma\left(\frac{d+1}{2}\right)\sqrt{2\lambda_1}}\epsilon + 1}} - c_3\sqrt{d}\right)^2} \\
&\leq 2e^{-c_4\left(\sqrt{N} - \sqrt{\frac{N-d-2}{\frac{\lambda_d}{d\sqrt{2\lambda_1}}\epsilon + 1}} - c_3\sqrt{d}\right)^2}, \qquad \text{(H.2)}
\end{aligned}
$$

for $\sqrt{N} > \sqrt{\frac{N-d-2}{\frac{\lambda_d}{d\sqrt{2\lambda_1}}\epsilon+1}} + c_3\sqrt{d}$ where $c_3, c_4 > 0$ are constants that do not depend on d or N. □

Appendix I. Choosing trade-off variables

We have

$$\begin{aligned}
\Pr\left(\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| \geq \epsilon\right) &\leq 8e^{-c_4\left(\sqrt{N}-\sqrt{\frac{N-d-2}{\frac{\lambda_d}{6d\sqrt{2\lambda_1}}\epsilon+1}}-c_3\sqrt{d}\right)^2} + 2^{N+2}e^{-\frac{\epsilon^2 M\lambda_d}{4608\delta_1\delta_2}} \\
&+ 4N\left(\frac{\delta_2}{d}e^{1-\frac{\delta_2}{d}}\right)^{d/2} + 8e^{\log N-\frac{M\delta_1}{2N}-o\left(\frac{M\delta_1}{2N}\right)} + 2^{N+2}e^{-\frac{\epsilon^2 M\lambda_d}{1152\delta_0}} \\
&+ 4N\left(\frac{\epsilon^2\lambda_d}{576d\delta_3}e^{1-\frac{\epsilon^2\lambda_d}{576d\delta_3}}\right)^{d/2} + 8Ne^{-\frac{N\delta_3}{2}} + 4e^{-\frac{1}{4}\left(\sqrt{\frac{\epsilon^2 N\lambda_d}{c_2}-d}-\sqrt{d}\right)^2} \\
&+ 4N\left(\frac{\delta_0}{d}e^{1-\frac{\delta_0}{d}}\right)^{d/2} + 4N\left(\frac{\epsilon^2\lambda_d}{2304d\delta_1}e^{1-\frac{\epsilon^2\lambda_d}{2304d\delta_1}}\right)^{d/2}, \qquad \text{(I.1)}
\end{aligned}$$

for $\sqrt{N} > \sqrt{\frac{N-d-2}{\frac{\lambda_d}{6d\sqrt{2\lambda_1}}\epsilon+1}} + c_3\sqrt{d}$ and $0 < \epsilon < 1$ where $c_1, c_2, c_3, c_4 > 0$ are absolute constants. The bounds we use require $\delta_2 > d$, $\delta_1 < \epsilon^2\lambda_d/2304d$, $\delta_1 > 2N\log(N)/M$, $\delta_0 > d$, $\delta_3 < \epsilon^2\lambda_d/576d$ and $N > dc_2/\epsilon^2\lambda_d$. Terms δ_0, δ_1, δ_2 and δ_3 need to be defined as functions of N, M and d such that the conditions arising from the tail bounds hold and the exponential terms' limit are 0 as $N, M \to \infty$. In order to achieve this, we start dealing with δ_0 first. The condition δ_0 needs to satisfy is

$$d(1+\log(\delta_0/d)) + 2\log N < \delta_0 < \frac{\epsilon^2 M\lambda_d}{1152N\log 2}.$$

The condition for δ_1 is

$$\frac{2N\log N}{M} < \delta_1 < \frac{\epsilon^2\lambda_d}{2304\left[d\left(\log\left(\frac{\epsilon^2\lambda_d}{2304d\delta_1}\right)+1\right)+2\log N\right]}.$$

The condition for δ_2 is

$$d(1+\log(\delta_2/d)) + 2\log N < \delta_2 < \frac{\epsilon^2 M\lambda_d}{4608N\delta_1\log 2}.$$

The condition for δ_3 is

$$\frac{2\log N}{N} < \delta_3 < \frac{\epsilon^2\lambda_d}{576\left[2\log N + d\left(\log\left(\frac{\epsilon^2\lambda_d}{576d\delta_3}\right)+1\right)\right]}.$$

We let $M = O\left(\frac{dN\log^3 N}{\lambda_d \epsilon^2}\right)$ together with, $\delta_0 = d\log^2 N, \delta_1 = \frac{4\lambda_d\epsilon^2}{d\log^2 N}, \delta_2 = d\log^2 N, \delta_3 = \frac{\epsilon^2\lambda_d}{1152 d\log N}$. Substituting these quantities in (I.1) gives,

$$\Pr\left(\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| \geq \epsilon\right) \leq 8e^{-c_4\left(\sqrt{N}-\sqrt{\frac{N-d-2}{\frac{\lambda_d}{6d\sqrt{2\lambda_1}}\epsilon+1}}-c_3\sqrt{d}\right)^2} + 2^{N+2}e^{-\frac{N\log^3 N}{18432\lambda_d\epsilon^2}}$$
$$+ 4N(\log^2 Ne^{1-\log^2 N})^{d/2} + 8e^{-4N(\log N-1)} + 2^{N+2}e^{-\frac{N\log N}{1152}}$$
$$+ 4N(2\log Ne^{1-2\log N})^{d/2} + 8Ne^{-\frac{N\epsilon^2\lambda_d}{2304 d\log N}} + 4N(\log^2 Ne^{1-\log^2 N})^{d/2}$$
$$+ 4e^{-\frac{1}{4}\left(\sqrt{\frac{\epsilon^2 N\lambda_d}{c_2}-d}-\sqrt{d}\right)^2} + 4N\left(\frac{\log^2 N}{9216}e^{1-\frac{\log^2 N}{9216}}\right)^{d/2}, \tag{I.2}$$

for $\sqrt{N} > \sqrt{\frac{N-d-2}{\frac{\lambda_d}{6d\sqrt{2\lambda_1}}\epsilon+1}} + c_3\sqrt{d}$ where $c_1, c_2, c_3, c_4 > 0$ are constants, $N > dc_2/\epsilon^2\lambda_d$, $N > \frac{2304 d\log^2 N}{\epsilon^2\lambda_d}$, $\log N > 347$, $\log N > 18(\lambda_d\epsilon^2)^{1/3}$. We can simplify this bound for large enough N. We consider terms with $\log^2 N$ and $1 - 2\log N$ first. One of the $\log^2 N$ terms is divided by 9216. When $\log^2 N$ is higher than this number, it will be dominated by the $2\log N$ term. Therefore, for $\log N > c_5$ where $c_5 = 18432$ we have,

$$\Pr\left(\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| \geq \epsilon\right) \leq 8e^{-c_4\left(\sqrt{N}-\sqrt{\frac{N-d-2}{\frac{\lambda_d}{6d\sqrt{2\lambda_1}}\epsilon+1}}-c_3\sqrt{d}\right)^2} + 2^{N+2}e^{-\frac{N\log^3 N}{18432\lambda_d\epsilon^2}}$$
$$+ 8e^{-4N(\log N-1)} + 16N(2\log Ne^{1-2\log N})^{d/2} + 8Ne^{-\frac{N\epsilon^2\lambda_d}{2304 d\log N}}$$
$$+ 4e^{-\frac{1}{4}\left(\sqrt{\frac{\epsilon^2 N\lambda_d}{c_2}-d}-\sqrt{d}\right)^2} + 2^{N+2}e^{-\frac{N\log N}{1152}}. \tag{I.3}$$

Now we consider the terms with the exponent $N\log^3 N$ and $N\log N$. For $\log N > 4\epsilon\sqrt{\lambda_d}$, we can reduce the bound to

$$\Pr\left(\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| \geq \epsilon\right) \leq 8e^{-c_4\left(\sqrt{N}-\sqrt{\frac{N-d-2}{\frac{\lambda_d}{6d\sqrt{2\lambda_1}}\epsilon+1}}-c_3\sqrt{d}\right)^2} + 8Ne^{-\frac{N\epsilon^2\lambda_d}{2304 d\log N}}$$
$$+ 16N(2\log Ne^{1-2\log N})^{d/2} + 2^{N+3}e^{-\frac{N\log N}{1152}} + 4e^{-\frac{1}{4}\left(\sqrt{\frac{\epsilon^2 N\lambda_d}{c_2}-d}-\sqrt{d}\right)^2}, \tag{I.4}$$

for $\sqrt{N} > \sqrt{\frac{N-d-2}{\frac{\lambda_d}{6d\sqrt{2\lambda_1}}\epsilon+1}} + c_3\sqrt{d}$, $N > dc_2/\epsilon^2\lambda_d$, $N > \frac{2304 d\log^2 N}{\epsilon^2\lambda_d}$, $\log N > 18(\lambda_d\epsilon^2)^{1/3}$, $\log N > c_5$, $\log N > 4\epsilon\sqrt{\lambda_d}$ where $c_1, c_2, c_3, c_4, c_5 > 0$ are absolute constants. Note that under given conditions, we have 2 terms that are competing, i.e. the term with $\log N$ and the term with $N/\log N$. Therefore, the bound reduces to

$$\Pr\left(\left\|\hat{\boldsymbol{\beta}} - c_1\boldsymbol{\beta}\right\| \geq \epsilon\right) \leq c_6 N\max\left\{\left(\frac{\sqrt{6\log N}}{N}\right)^d, e^{-\frac{N\epsilon^2\lambda_d}{c_7 d\log N}}\right\}, \tag{I.5}$$

for $N > \frac{c_8 d\log^2 N}{\epsilon^2\lambda_d}$ and $0 < \epsilon < 1$ where $c_1, c_6, c_7, c_8 > 0$ are absolute constants.