MINIMAX SUPERVISED CLUSTERING IN THE ANISOTROPIC GAUSSIAN MIXTURE MODEL: A NEW TAKE ON ROBUST INTERPOLATION

By Yiqiu Shen*, Stanislav Minsker* and Mohamed Ndaoud*

*University of Southern California**

We study the supervised clustering problem under the two-component anisotropic Gaussian mixture model in high dimensions in the non-asymptotic setting. We first derive a lower and a matching upper bound for the minimax risk of clustering in this framework. We also show that in the high-dimensional regime, the linear discriminant analysis (LDA) classifier turns out to be sub-optimal in a minimax sense. Next, we characterize precisely the risk of regularized supervised least squares classifiers under ℓ_2 regularization. We deduce the fact that the interpolating solution (0 training error solution) may outperform the regularized classifier, under mild assumptions on the covariance structure of the noise. Our analysis also shows that interpolation can be robust to corruption in the covariance of the noise when the signal is aligned with the "clean" part of the covariance, for the properly defined notion of alignment. To the best of our knowledge, this peculiar phenomenon has not yet been investigated in the rapidly growing literature related to interpolation. We conclude that interpolation is not only benign but can also be optimal and in some cases robust.

1. Introduction. The topic of overparametrization has gain tremendous interest in recent literature devoted to high dimensional statistics. Previously, it was widely believed that regularization yields the best generalization power because of bias-variance tradeoff. Recently, it was discovered that interpolation also generalizes well when number of covariates exceeds sample size. This phenomenon, termed "benign overfitting" by [2], has been intensively investigated in the regression setting. In this work, we study the problem of clustering. In particular, we derive the bounds for the generalization error under different settings. The model we consider is a binary sub-Gaussian mixture model with unknown, anisotropic noise.

1.1. *Statement of the problem.* Consider the simple two component Gaussian mixture model, where we are given

$$Y = \theta\eta^\top + W$$

or equivalently for all $i = 1, \dots, n$

$$\eta_i Y_i = \theta + \eta_i W_i$$

where $\theta \in \mathbb{R}^p$ is a center vector, $\eta \in \{-1, 1\}^n$ a label vector and W a random matrix with i.i.d vectors $W_i \sim \mathcal{N}(0, \Sigma)$ (or sub-Gaussian) and Σ is full rank. We mostly focus on the supervised setting, where we are given a classifier $\hat{\eta}$ and a new independent observation (Y_{n+1}, η_{n+1}) such that η_{n+1} is a Rademacher random variable. We want to analyze its generalization error given by

$$R_\Sigma(\hat{\eta}) := \mathbb{P}\left(\hat{\eta}(Y_{n+1}) \neq \eta_{n+1}\right),$$

where $\mathbb{P}$ is the probability under the model described above. When there is no ambiguity we will omit the subscript Σ from R_Σ. In particular we want to analyze the minimax risk

$$\inf_{\hat{\eta}} \sup_{\|\theta\| \geq \Delta} R(\hat{\eta}),$$

and study necessary and sufficient conditions on Δ for consistent clustering i.e. conditions on $(\Delta_n)_n$ such that

$$\inf_{\hat{\eta}} \sup_{\|\theta\| \geq \Delta_n} R(\hat{\eta}) \underset{n \to \infty}{\to} 0.$$

The case $\Sigma = I_p$ was studied in [9]. In particular it is shown there that

$$\inf_{\hat{\eta}} \sup_{\|\theta\| \geq \Delta} R(\hat{\eta}) \approx \exp\left(-(1 + o_n(1)) \frac{\Delta^4}{2(\Delta^2 + \frac{p}{n})}\right).$$

The case of general but known Σ was studied in [12]. Following similar arguments as in [9] we can actually show in the general case that

$$\inf_{\hat{\eta}} \sup_{\|\theta\|_\Sigma \geq \Delta} R(\hat{\eta}) \approx \exp\left(-(1 + o_n(1)) \frac{\Delta^4}{2(\Delta^2 + \frac{p}{n})}\right),$$

where $\|\theta\|_{\Sigma^{-1}}^2 = \theta^\top \Sigma^{-1} \theta$. In particular the condition $\|\theta\|_\Sigma^2 \gg \sqrt{p/n} + 1$ is necessary and sufficient for consistency under the norm $\|.\|_\Sigma$. Moreover a minimax optimal classifier is the LDA classifier given by

$$\hat{\eta}_{\text{LDA}} = \text{sign}\left(\left\langle \Sigma^{-1} \sum_{i=1}^n Y_i \eta_i, Y_{n+1} \right\rangle\right),$$

(cf. [12] and references therein and the recent work of [5]. All these works consider anisotropic mixtures in the low dimensional case). In this project we are interested in the case of unknown Σ in high dimensions (i.e $p \gg n$) where estimation of both θ and Σ becomes challenging.

1.2. *Related work.* [2] introduce the concept of benign overfitting, a phenomenon that interpolation can give accurate prediction in linear regression. This is counter intuitive compared with the traditional thinking that overfitting leads to severe bias-variance tradeoff. This work provides a complete proof of the upperbound of excess risk of min norm estimator. [13] consider the case where noise of the linear model is anisotropic. They propose the idea of aligned and misaligned prior on true coefficients and data covariance and analyse the asymptotic behavior of the prediction risk of the generalized ridge regression estimator in the overparameterized regime. There are a lot of works considering overparametrized regression, but our focus is to derive bounds for misclassification rate for clustering (sub)Gaussian mixture models. We write our proofs in the style of [9]. Previously, [11] explore the isotropic case where SVM solution linear interpolates the data. They argue that when p is large enough, error of SVM interpolator (which is equivalently least square solution) goes to 0, under different conditions both in high SNR regime and low SNR regime (the bi-level ensemble is less general than our results). [12] and [5] consider anisotropic clustering, but not in overparametrization sense (they are more aligned with [9]). Both propose efficient algorithms. The former one is via uncoupled regression and perturbed gradient descent while the latter one is based on Lloyd's algorithm. [7] provides bound for misclassification error in a more general structure setting via analyzing properties of RKHS. [3] is very close to our work, but only considers the case of $r(\Sigma) \geq n$ and no regularization is considered.

Reference	Type	Noise	$p \gg n$	Asymptotic
[2]	Regression	Isotropic	Yes	No
[13]	Regression	Anisotropic	Yes	Yes
[9]	Classification	Isotropic	Yes	No
[11]	Classification	Anisotropic	Yes	No
[12]	Classification	Anisotropic	No	No
[7]	Classification	Isotropic	Yes	No
[5]	Classification	Anisotropic	No	Yes
[4]	Classification	Anisotropic	Yes	No
[8]	Classification	Anisotropic	Yes	Yes
[3]	Classification	Anisotropic	Yes	No
Our Work	Classification	Anisotropic	Yes	No

Table 1: Summary of the related work.

1.3. *Contribution.* The closest works to our paper are [11, 3]. Both works consider the case where $r(\Sigma) \geq n$. We summarize below their findings: Our

	Wang and Thrampoulidis [11]	Cao et al.[3]
Proliferation	$\mathrm{Tr}(\Sigma) > C\left(\|\mathbf{\Sigma}\|_F \cdot n\sqrt{\log n} + \|\Sigma\|_\infty \cdot n\sqrt{n}\log n + 1\right)$ $\mathrm{Tr}(\Sigma) > C_1 n\sqrt{\log(2n)}\|\theta\|_\Sigma$	$\mathrm{Tr}(\Sigma) \geq C\max\left\{n^{3/2}\|\Sigma\|_\infty, n\|\Sigma\|_F\right\}$ $\mathrm{Tr}(\Sigma) > C_1 n\sqrt{\log(n)} \cdot \|\theta\|_\Sigma$
Error bound	$\exp\left(\frac{-\left(\|\theta\|_2^2 - \frac{C_1 n\|\theta\|_\Sigma^2}{\mathrm{Tr}(\Sigma)} - C_2\|\theta\|_\Sigma\right)^2}{C_3\max\left\{1, \frac{n^2\|\theta\|_\Sigma^2}{\mathrm{Tr}(\Sigma)^2}\right\}\|\Sigma\|_F^2 + C_4\|\theta\|_\Sigma^2}\right)$	$\exp\left(\frac{-C'\|\theta\|_2^4}{\|\theta\|_\Sigma^2 + \|\Sigma\|_F^2/n + \|\Sigma\|_\infty^2}\right)$

contributions are three folds:

- First, we derive the minimax generalization risk for clustering in the anisotropic sub-Gaussian model and show that the averaging classifier that is adaptive is also minimax optimal.
- Next, we show that proliferation happens under the mild conditions: $\mathrm{Tr}(\Sigma) \geq Cn\log(n)\|\Sigma\|_\infty$ and $\mathrm{Tr}(\Sigma) > C_1 n\sqrt{\log(n)} \cdot \|\theta\|_\Sigma$. These condition are strictly better than previous ones and hold in particular when $r(\Sigma) \geq n\log(n)$.
- Finally we show that under mild assumptions the interpolating solution is minimax optimal when $r(\Sigma) \geq Cn$ leading to a better bound to previous works. We also show that, under corruption of the covariance matrix, interpolation can lead to a robust classifier, a feature that is not available for the averaging oracle. Hence not only interpolation can be benign, but can also be optimal and robust!

2. Minimax clustering: the supervised case. In this section we consider the supervised clustering problem when Σ is not necessarily known. For a matrix A we define its effective rank by $r(A) := \mathrm{Tr}(A)/\|A\|_\infty$. We first establish a lower bound result.

THEOREM 2.1. *Let $\Delta, \lambda, r > 0$. Then*

$$\inf_{\hat{\eta}} \sup_{r(\Sigma^2)=r, \|\Sigma\|_\infty=\lambda} \sup_{\|\theta\|^2 \geq \Delta^2\lambda} R(\hat{\eta}) \geq C\exp\left(-c\frac{\Delta^4}{\Delta^2 + \frac{r}{n}}\right),$$

for some $c, C > 0$ where the infimum is over all measurable classifiers $\hat{\eta}(Y)$.

Notice that the norm of θ is the Euclidean norm and not the Mahalanobis one that would lead to a different lower bound. The proof of the lower bound is inspired from the lower bound proof in [9] that only holds for isotropic noise. As for the upper bound, we show that the averaging linear classifier

$$\hat{\eta}_{\mathrm{ave}} = \mathrm{sign}\left(\left\langle \sum_{i=1}^n Y_i\eta_i, Y_{n+1}\right\rangle\right),$$

is minimax optimal. In fact we can show the following result.

THEOREM 2.2. *Let $\Delta > 0$. For any covariance matrix Σ we have*

$$\sup_{\|\theta\|^2 \geq \Delta^2 \|\Sigma\|_\infty} R(\hat{\eta}_{ave}) \leq C \exp\left(-c \frac{\Delta^4}{\Delta^2 + \frac{r(\Sigma^2)}{n}}\right),$$

for some $c, C > 0$.

Theorem 2.2 provides a matching upper bound to Theorem 2.1. In particular this implies that consistency, under the Euclidean norm, happens under the condition $\|\theta\|^2 \gg \|\Sigma\|_\infty(\sqrt{r(\Sigma^2)/n} + 1)$. Among other conclusions this implies for instance that, from a minimax perspective, the averaging classifier outperforms the LDA one. This phenomenon is only possible in high dimensions. It is easy actually to show that, when $p \ll n$, then LDA outperforms the averaging classifier for any given center θ. Notice that the last statement is stronger than a minimax comparison.

REMARK 2.1. *Minimax clustering: the unsupervised case. We can actually extend the result of this section to the unsupervised case. Using the same procedure as in [9] we claim the existence of a polynomial time method that is minimax optimal for clustering.*

3. Interpolation vs Regularization in Gaussian mixtures.

In this section we study the risk of the regularized OLS estimators. While it is more common to study SVM for classification, recent works [1, 6] have shown that in high dimensions ($p = \Omega(n \log n)$), both SVM and OLS solutions coincide under mild conditions. This phenomenon is also known in the literature as proliferation of support vectors. Hence is high dimensions, it is sufficient to study the least squares estimator and then show that it coincide with the hard-margin SVM solution. For the rest of this section, our goal is the study the risk of the following family of supervised estimators solving

$$\min_{\hat{\theta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (\eta_i - \langle Y_i, \hat{\theta} \rangle)^2 + \lambda \|\hat{\theta}\|^2.$$

Observe that the case $\lambda = 0$ and $p \geq n$ leads to interpolation and more precisely to the minimum ℓ_2-norm interpolating solution (cf [2] for interpolation in regression and [11, 3] for interpolation in clustering of Gaussian *mixtures, cf. also [7]). For each $\lambda > 0$, the corresponding estimator $\hat{\theta}_\lambda$ is proportional to*

$$\hat{\theta}_\lambda = \frac{1}{n}\left(\lambda I_p + \frac{1}{n}YY^\top\right)^{-1} Y\eta = \frac{1}{n}Y\left(\lambda I_n + \frac{1}{n}Y^\top Y\right)^{-1}\eta.$$

Each estimator $\hat{\theta}_\lambda$ leads to a linear classifier defined through

$$\hat{\eta}_\lambda(.) = \operatorname{sign}\left(\left\langle \hat{\theta}_\lambda, .\right\rangle\right).$$

In what follows we compute $R(\hat{\eta}_\lambda)$. We also provide sufficient conditions for the matrix Σ such that the interpolating classifier (corresponding to $\lambda = 0$) has at least the same performance as the oracle (or the averaging classifier).

Notice that as λ goes to ∞ we recover the averaging classifier. We emphasize here that if we replace the regularizing term $\|\theta\|^2$ by $\|\theta\|_\Sigma^2$ (not adaptive to Σ), then our claims may no longer be true.

As a side note, We suspect that the excess risk of estimation θ gets smaller for large values of λ although this is not necessarily the case for the classification risk. This suggests that the excess risk of estimation is not the right metric to evaluate classification performance.

In order to state our main result, we define the following quantities. For any integer k, let

$$r_k(\Sigma) = \frac{\sum_{i=k+1}^{p} \lambda_i}{\lambda_{k+1}},$$

where $(\lambda_k)_k$ is the decreasing sequence of eigenvalues of Σ. For a given covariance matrix Σ, we define k_Σ^* such that

$$k_\Sigma^*(\lambda) = \min\left\{k \geq 0, r_k(\Sigma) + \frac{\lambda}{\lambda_{k+1}} \geq C_1 n\right\},$$

for some constant $C_1 > 1$ and $k_\Sigma^*(\lambda) = p+1$ if the above set if empty. The reader may observe that $k_\Sigma^*(.)$ is decreasing with λ and that $k_\Sigma^*(\lambda) = 0$ if $r(\Sigma) \geq C_1 n$. In what follows we will require k_Σ^* to be smaller than $n/2$ which means that we are not allowing more than a fraction of n eigenvalues to be much larger than the remainder of the spectrum. This is a large class of covariance matrices, and contain in particular the case $\Sigma = I_p + R$ where R is a low rank perturbation/corruption of the isotropic noise.

THEOREM 3.1. *Let $\Delta > 0, \lambda \geq 0$. Assume that $k_\Sigma^*(\lambda) \leq n/2$ and that $\|\theta\|_\Sigma^2/\|\theta\|^2 \leq C(\sum_{i>k^*} \lambda_i/n + \lambda)$. Then for some constants $c, C > 0$ we have* *with probability* $1-\delta-e^{-cn}$ *that*

$$R(\hat{\eta}_\lambda) \le C\exp\left(-c\frac{\|\theta\|^4}{\theta^T\Sigma\theta(1+k^*)+\frac{k^*\lambda_{k^*}^2+\sum_{i>k^*}\lambda_i^2(\Sigma)}{n}+\frac{(k^*\lambda_{k^*}^2+\lambda_{k^*+1}^2)\log(1/\delta)}{n}}\right),$$

where $k^*=k^*_\Sigma(\lambda)$.

The condition on the alignment of θ *with* Σ *can be understood as follows. Remember that we are thinking of the first* k^* *eigenvectors of* Σ *as outliers. Hence as long as* $\|\theta\|_\Sigma^2/\|\theta\|^2 \le \lambda_{k^*+1}$ *then* $\|\theta\|_\Sigma^2/\|\theta\|^2 \le C(\sum_{i>k^*}\lambda_i/n+\lambda)$ *holds. Hence the latter condition means simply that the vector* θ *is only allowed to be aligned with the "clean" part of the covariance* Σ.

When $k^*=0$ *(or equivalently* $r(\Sigma)\ge C_1 n$*) we recover the bound in [3] by taking* $\delta=e^{-cn}$. *Our result is stronger since we show that the bound holds with probability* $1-e^{-cn}$ *while they only show that the same bound holds with probability* $1-1/n$. *Moreover, under the mild condition* $r(\Sigma^2)\ge\log(n)$, *then by taking* $\delta=1/n$ *we show that with probability* $1-1/n$ *we have*

$$R(\hat{\eta}_\lambda) \le C\exp\left(-c\frac{\|\theta\|^4}{\theta^T\Sigma\theta+\frac{\mathrm{Tr}(\Sigma^2)}{n}}\right).$$

Hence interpolation leads to the same bound as the averaging oracle in this case. Without any further assumptions on Σ *we get the following minimax result*

COROLLARY 3.1. *Let* $\Delta>0,\lambda\ge0$. *Assume that* $r(\Sigma)\ge C_1 n$. *Then for some constants* $c,C>0$ *we have that*

$$\sup_{\|\theta\|\ge\Delta} R(\hat{\eta}_\lambda) \le C\exp\left(-c\frac{\Delta^4}{\Delta^2\|\Sigma\|_\infty+\frac{\mathrm{Tr}(\Sigma^2)}{n}}\right)+e^{-cn}.$$

The above result suggests that under a mild condition on the noise covariance, not only interpolation is benign but it is also optimal in a minimax sense. This also means that interpolation is better for classification than in regression ([2]) since it does not suffer from a bias term which leads to a bad worst-case performance.

Unlike previous works [3, 11], our result is more geenral since we allow k^* *to be non zero. It is straightforward to observe that* $k^*\lambda_{k^*}^2+\sum_{i>k^*}\lambda_i^2(\Sigma)$ *is increasing with* λ. *The latter quantity could be seen as a truncated trace of* Σ^2 *where we truncate the large eigenvalues of* Σ. *The bound in Theorem 3.1, in particular, gets smaller as* λ *goes to* 0 *which suggests that interpolation may* *outperform regularization in some cases, especially under several corruption where a finite number of eigenvalues are much larger than the rest of the spectrum.*

Remember that all our results require $k^* \leq n/2$. We may wonder here what happens if such k^* is much larger than n and our simulations (see further) suggest that interpolation behaves poorly in this case.

4. Proliferation of support vectors in high dimensions under the subGaussian mixture model. *In this section, we provide sufficient conditions for proliferation of support vectors. Based on [6], both $\hat{\theta}_{SVM}$ and $\hat{\theta}_0$ coincide if and only if*

$$\forall i = 1, \dots, n \quad \eta_i e_i^\top (Y^\top Y)^{-1} \eta > 0,$$

where $(e_i)_{i=1,\dots,n}$ is the Euclidean canonical basis. The main result is stated next.

THEOREM 4.1. *Assume that $k^* \log^2(n) \leq Cn$, $\sum_{i>k} \lambda_i^2 n \log(n) \leq C(\sum_{i>k} \lambda_i)^2$ and $\|\theta\|_\Sigma \sqrt{(1+k^*)\log(n)} \leq C \sum_{i>k} \lambda_i / n$, then with probability $1 - 1/n$ we have*

$$\hat{\theta}_{SVM} = \hat{\theta}_0.$$

When $k^* = 0$, the sufficient conditions can read as:

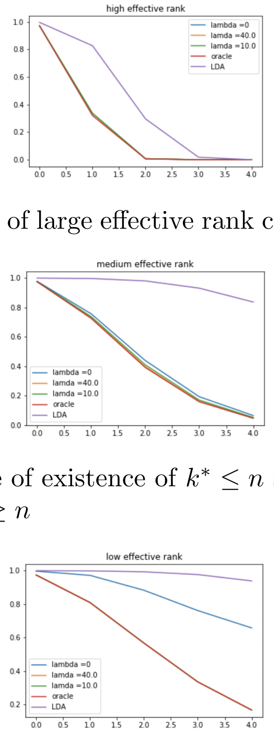
- $\|\theta\|_\Sigma \leq C \operatorname{Tr}(\Sigma)/n$.
- $\operatorname{Tr}(\Sigma^2) n \log(n) \leq C(\operatorname{Tr}(\Sigma))^2$.

The first condition (signal dependent) is also required in both papers [3, 11]. As for the dimension dependent condition, our requirement is much milder than the was proposed in both previous papers. To compare these results, we can think of the case $\Sigma = I_p$, where our condition reads as $p = \Omega(n \log(n))$ while previous require $p = \Omega(n^{3/2} \log(n))$. Our result also suggests that $r(\Sigma) = \Omega(n \log(n))$ is sufficient for proliferation under the sub-Gaussian mixture model which confirms the general conjecture stated by [6].

5. Numerical experiments. *In this section, we propose some numerical experiments to endorse our theory. We consider the worst-case generalization error for three cases of the noise covariance matrix.*

- *The case of large effective rank corresponds to $k^* = 0$.*
- *The case of medium effective rank corresponds to $k^* = n/4$.*
- *The case of small effective rank corresponds to $k^* \gg n$.*

Our simulations suggest that interpolation has a similar performance to the oracle as long as k^* is smaller than n. We can also see that interpolation performs poorly, in a minimax sense, compared to other methods as k^* grows.



(a) Case of large effective rank covariance

(b) Case of existence of $k^* \leq n$ such that $r_{k^*}(\Sigma) \geq n$

(c) Case of small effective rank $r_k(\Sigma)$ for all k

Fig 1: Comparison of the worst-case generalization errors of some supervised learners under different covariance scenarios.

A recent paper ([13]) considers interpolation as a special case of ridge regression in the case where both design and signal are anisotropic. There are some nice ideas that we can definitely use in the setup of Gaussian mixture. In particular they introduce a notion of misalignment between signal and the covariance that gives intuition about when the interpolation is better than regularization. In our case, the worst case scenario happens when θ corresponds to the top eigenvector of Σ (case of alignment). It turns out that the picture is completely different in the opposite scenario where θ cor-

responds to the smallest eigenvector (case of misalignment) when $r(\Sigma) \ll n$*. In this case, simulations show that interpolation has a closer performance to LDA and they both beat regularized classifiers (and the oracle). We may actually want to prove a theoretical result in this setup. This would imply that interpolation can actually be better than regularization in some cases.*

Acknowledgements. *Authors acknowledge support by the National Science Foundation grant CCF-1908905.*

References.

[1] Ardeshir, N., Sanford, C. *and* Hsu, D. *(2021). Support vector machines and linear regression coincide with very high-dimensional features.*

[2] Bartlett, P. L., Long, P. M., Lugosi, G. *and* Tsigler, A. *(2020). Benign overfitting in linear regression. Proceedings of the National Academy of Sciences.*

[3] Cao, Y., Gu, Q. *and* Belkin, M. *(2021). Risk Bounds for Over-parameterized Maximum Margin Classification on Sub-Gaussian Mixtures.*

[4] Chatterji, N. S. *and* Long, P. M. *(2020). Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. arXiv preprint arXiv:2004.12019.*

[5] Chen, X. *and* Zhang, A. Y. *(2021). Optimal Clustering in Anisotropic Gaussian Mixture Models. arXiv preprint arXiv:2101.05402.*

[6] Hsu, D., Muthukumar, V. *and* Xu, J. *(2020). On the proliferation of support vectors in high dimensions.*

[7] Liang, T. *and* Recht, B. *(2021). Interpolating Classifiers Make Few Mistakes.*

[8] Mai, X. *and* Liao, Z. *(2019). High dimensional classification via empirical risk minimization: Improvements and optimality. arXiv preprint arXiv:1905.13742.*

[9] Ndaoud, M. *(2018). Sharp optimal recovery in the two component gaussian mixture model. arXiv preprint arXiv:1812.08078.*

[10] Rudelson, M. *and* Vershynin, R. *(2013). Hanson-Wright inequality and subgaussian concentration.*

[11] Wang, K. *and* Thrampoulidis, C. *(2020). Benign Overfitting in Binary Classification of Gaussian Mixtures. arXiv preprint arXiv:2011.09148.*

[12] Wang, K., Yan, Y. *and* Diaz, M. *(2020). Efficient Clustering for Stretched Mixtures: Landscape and Optimality. arXiv preprint arXiv:2003.09960.*

[13] Wu, D. *and* Xu, J. *(2020). On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression. arXiv preprint arXiv:2006.05800.*

APPENDIX A: PROOFS OF MINIMAX CLUSTERING

A.1. Proof of Theorem 2.1. *Fix $\lambda, r > 0$ and let Σ be a diagonal PSD matrix such that $r(\Sigma^2) = r$ and $\|\Sigma\|_\infty = \lambda$. Then*

$$\inf_{\hat{\eta}} \sup_{r(\hat{\Sigma}^2)=r, \|\hat{\Sigma}\|_\infty=\lambda} \sup_{\|\theta\|^2 \geq \Delta^2\lambda} R_{\hat{\Sigma}}(\hat{\eta}) \geq \inf_{\hat{\eta}} \sup_{\|\theta\|^2 \geq \Delta^2\lambda} R_{\Sigma}(\hat{\eta}).$$

We can simply focus on showing the result for the above diagonal matrix Σ. The proof is decomposed in two steps.

- ***A dimension independent lower bound:***
  We have that

$$2 \inf_{\hat{\eta}} \sup_{\|\theta\|^2 \geq \Delta^2\lambda} R_{\Sigma}(\hat{\eta}) \geq \inf_{\tilde{\eta}} \mathbb{E}_\pi \mathbf{E}_{(\bar{\theta},\eta)} |\tilde{\eta}(Y, Y_{n+1}) - \eta_{n+1}|$$

for any prior π on (θ, η) such that $\|\theta\|^2 \geq \Delta^2\lambda$. Let $\bar{\theta}$ be a vector in $\mathbf{R}^p$ such that $\|\bar{\theta}\|^2 = \Delta^2\lambda$. Placing an independent Rademacher prior π on η_{n+1}, and fixing θ, it follows that

$$\inf_{\tilde{\eta}} \mathbb{E}_\pi \mathbf{E}_{(\bar{\theta},\eta)} |\tilde{\eta}(Y, Y_{n+1}) - \eta_{n+1}| \geq \inf_{\bar{\eta}} \mathbb{E}_\pi \mathbf{E}_{(\bar{\theta},\eta)} |\bar{\eta}(Y_{n+1}) - \eta_{n+1}|, \tag{1}$$

where $\bar{\eta}(Y_{n+1}) = \mathbf{E}(\tilde{\eta}(Y, Y_{n+1})|Y) \in [-1, 1]$. The last inequality holds because of independence between Y and Y_{n+1}. We define, for $\epsilon \in \{-1, 1\}$, $\tilde{f}_\epsilon(.)$ the density of the observation Y_{n+1} conditionally on the value of $\eta_{n+1} = \epsilon$. Now, using Neyman-Pearson lemma and the explicit form of $\tilde{f}_\epsilon$, we get that the selector η^* given by

$$\eta^* = \operatorname{sign}\left(\bar{\theta}^\top \Sigma^{-1} Y_{n+1}\right),$$

is the optimal selector that achieves the minimum of the RHS of (1). To show that, we remind the reader that the distribution of $Y_{n+1} = \eta_{n+1}\bar{\theta} + W_{n+1}$ is given by $\mathscr{N}(\eta_{n+1}\bar{\theta}, \Sigma)$. Hence

$$\tilde{f}_\epsilon(Y_{n+1}) = (2\pi)^{-p/2}|\Sigma|^{-1/2} e^{-\frac{1}{2}(Y_{n+1}-\epsilon\bar{\theta})^T \Sigma^{-1}(Y_{n+1}-\epsilon\bar{\theta})}.$$

It follows that

$$\begin{aligned}\frac{\tilde{f}_1(Y_{n+1})}{\tilde{f}_{-1}(Y_{n+1})} &= \frac{(2\pi)^{-p/2}|\Sigma|^{-1/2} e^{-\frac{1}{2}(Y_{n+1}-\bar{\theta})^T \Sigma^{-1}(Y_{n+1}-\bar{\theta})}}{(2\pi)^{-p/2}|\Sigma|^{-1/2} e^{-\frac{1}{2}(Y_{n+1}+\bar{\theta})^T \Sigma^{-1}(Y_{n+1}+\bar{\theta})}} \\ &= e^{2\bar{\theta}^T \Sigma^{-1} Y_{n+1}}\end{aligned}$$

By Neyman-Pearson lemma, we can now conclude that

$$\eta^* = \operatorname{sign}\left(\bar{\theta}^\top \Sigma^{-1} Y_{n+1}\right).$$

Plugging this value in (1), we know further that

$$\inf_{\bar{\eta}} \mathbf{E}_\pi |\bar{\eta}(Y_{n+1}) - \eta_{n+1}| = R(\eta^*).$$

It is now straightforward that

$$R(\eta^*) = \Phi^c\left(\sqrt{\bar{\theta}^\top \Sigma^{-1} \bar{\theta}}\right) \geq C e^{-c\bar{\theta}^\top \Sigma^{-1} \bar{\theta}},$$

for some $c, C > 0$. The above inequality holds for all $\bar{\theta}$ as long as $\|\bar{\theta}\|^2 = \Delta^2 \lambda$. The worst case is reached for $\bar{\theta}$ being co-linear with the top eigenvector of Σ since $\bar{\theta}^\top \Sigma^{-1} \bar{\theta} = \Delta^2$. Hence we get the lower bound

$$\inf_{\hat{\eta}} \sup_{\|\theta\|^2 \geq \Delta^2 \lambda} R_\Sigma(\hat{\eta}) \geq C e^{-c\Delta^2}.$$

In order to conclude we only need to derive the other lower bound

$$\inf_{\hat{\eta}} \sup_{\|\theta\|^2 \geq \Delta^2 \lambda} R_\Sigma(\hat{\eta}) \geq C e^{-cn\Delta^4/r}.$$

For the rest of the proof we only focus on the case $100 \leq \Delta^2 \leq 10r/n$, otherwise the dimension independent lower bound dominates.

- ***A dimension dependent lower bound:***
Since Σ is diagonal, we write $\Sigma = \operatorname{diag}(d_1, \dots, d_p)$ where $\lambda = d_1 \geq d_2 \geq \cdots \geq d_p > 0$. Using Theorem 1 in [9] we get

$$2 \inf_{\hat{\eta}} \sup_{\|\theta\|^2 \geq \Delta^2 \lambda} R_\Sigma(\hat{\eta}) \geq \inf_{T \in [-1,1]} \mathbb{E}_\pi \mathbf{E}_{(\theta,\eta)} |T(Y, Y_{n+1}) - \eta_{n+1}| - 2\pi(\|\theta\|^2 \leq \Delta^2 \lambda),$$

for any prior π on (θ, η). The second term in the above lower bound accounts for the constraint on θ. In what follows we choose π^D to be a product prior on $(\theta$,$\eta_{n+1})$ such that η_{n+1} is a Rademacher random variable and θ is an independent random vector such that $\theta \sim \mathscr{N}(0, D)$ where D is a diagonal matrix such that $D_{jj} = 2\Delta^2 \lambda \frac{d_j^2}{\sum_{i=1}^p d_i^2}$. Using the Hanson-Wright inequality it comes out that

$$\pi^D(\|\theta\|^2 \leq \Delta^2 \lambda) \leq C e^{-c\, r},$$

for some $c, C > 0$. Hence, and since $\Delta^2 \leq r/n$, we only need to show, for n large enough, that

$$\inf_{T \in [-1,1]} \mathbb{E}_\pi \mathbf{E}_{(\theta,\eta)} |T(Y, Y_{n+1}) - \eta_{n+1}| \geq C e^{-cn\Delta^4/r},$$

for some $c, C > 0$. We define, for $\epsilon \in \{-1, 1\}$, $\tilde{f}_\epsilon$ the density of the observation (Y, Y_{n+1}) given $\eta_{n+1} = \epsilon$. Using Neyman-Pearson lemma, we get that

$$\eta^{**} = \begin{cases} 1 & \textit{if } \tilde{f}_1(Y, Y_{n+1}) \geq \tilde{f}_{-1}(Y, Y_{n+1}), \\ -1 & \textit{else}, \end{cases}$$

minimizes $\mathbb{E}_{\pi^D} \mathbf{E}_{(\theta,\eta)} |T(Y, Y_{n+1}) - \eta_{n+1}|$ over all functions of (Y, Y_{n+1}) with values in $[-1, 1]$. Using the independence of the rows of Y we have

$$\tilde{f}_\epsilon(Y) = \prod_{j=1}^p \frac{e^{-\frac{1}{2} L_j^\top (\Sigma_\epsilon^j)^{-1} L_j}}{(2\pi)^{p/2} |\Sigma_\epsilon^j|},$$

where L_j is the j-th row of the matrix (Y, Y_{n+1}) and $\Sigma_\epsilon^j = d_j \mathbf{I}_{n+1} + D_{jj} \eta_\epsilon \eta_\epsilon^\top$. We denote by η_ϵ the binary vector such that $\eta_{n+1} = \epsilon$ and the other components are known. It is easy to check that $|\Sigma_\epsilon^j| = d_j + D_{jj}(n+1)$, hence it does not depend on ϵ. A simple calculation leads to

$$(\Sigma_\epsilon^j)^{-1} = (1/d_j) \mathbf{I}_n - \frac{D_{jj}/d_j^2}{1 + D_{jj} n / d_j} \eta_\epsilon \eta_\epsilon^\top$$

$$
\begin{aligned}
&= (1/d_i)\mathbf{I}_n - \frac{2\Delta^2\lambda/\sum d_i^2}{1+2n\Delta^2 d_j/\sum d_i^2}\eta_\epsilon\eta_\epsilon^\top \\
&= (1/d_i)\mathbf{I}_n - \frac{2\Delta^2\lambda/\sum d_i^2}{1+2(n\Delta^2 d_j)/(\lambda r)}\eta_\epsilon\eta_\epsilon^\top .
\end{aligned}
$$

Hence

$$
\begin{aligned}
\frac{\tilde f_1(Y)}{\tilde f_{-1}(Y)} &= \prod_{j=1}^p e^{-\frac{1}{2}L_j^\top((\Sigma_1^j)^{-1}-(\Sigma_{-1}^j)^{-1})L_j} \\
&= \prod_{j=1}^p \exp\left(\frac{2\Delta^2\lambda/\sum d_i^2}{1+(2n\Delta^2 d_j)/(\lambda r)}L_{j,n+1}\sum_{k=1}^n L_{jk}\eta_k\right) \\
&= \exp\left(\frac{2\Delta^2\lambda}{\sum d_i^2}\sum_{k=1}^n \eta_k \sum_{j=1}^p \frac{L_{jk}L_{j,n+1}}{1+(2n\Delta^2 d_j)/(\lambda r)}\right) \\
&= \exp\left(\frac{2\Delta^2\lambda}{\sum d_i^2}\langle Y_{n+1}, \sum_{k=1}^n \eta_k \tilde D Y_k\rangle\right).
\end{aligned}
$$

where $\tilde D = \operatorname{diag}\left(\frac{1}{1+(2n\Delta^2 d_i)/(\lambda r)}\right)_{i=1,\dots,p}$*. We conclude that the optimal selector is given by*

$$
\eta^{**} = \operatorname{sign}\left(Y_{n+1}^\top\left(\sum_{k=1}^n \eta_k \tilde D Y_k\right)\right)
$$

and that

$$
R(\eta^{**}) = \mathbb{P}((\tilde D Y^\top \eta)^\top Y_{n+1} < 0)
$$

Let us denote by $\hat\theta = \frac{1}{n}\sum_{i=1}^n Y_i\eta_i = \theta + \xi$ *where* $\xi = \frac{1}{n}\sum_{i=1}^n W_i\eta_i$*. Then*

$$
R(\eta^{**}) = \mathbf{E}\left(\Phi^c\left(\frac{\langle\theta, \tilde D\hat\theta\rangle}{\sqrt{\hat\theta^\top \tilde D\Sigma\tilde D\hat\theta}}\right)\right).
$$

Observing that the eigenvalues of $\tilde D$ *belong to* $[1/3, 1]$ *and that* $\tilde D\Sigma\tilde D \succeq \Sigma/9$ *in a PSD sense, it comes out that*

$$
R(\eta^{**}) \geq \mathbf{E}\left(\Phi^c\left(\frac{3\langle\theta, \tilde D\hat\theta\rangle}{\sqrt{\hat\theta^\top\Sigma\hat\theta}}\right)\right) \geq C\mathbf{E}\left(\exp\left(-c\frac{\|\theta\|^4 + \langle\theta, \tilde D\xi\rangle^2}{\hat\theta^\top\Sigma\hat\theta}\right)\right),
$$

for some $c, C > 0$*. It comes out that*

$$
R(\eta^{**}) \geq C\mathbf{E}\left(\exp\left(-c\frac{\|\theta\|^4 + \langle\theta, \tilde D\xi\rangle^2}{\xi^\top\Sigma\xi - 2\|\theta\|^2\lambda}\right)\right).
$$

Consider the three events

$$\mathbb{A}_1 = \{\|\theta\|^2 \leq 2\Delta^2\lambda\},$$

$$\mathbb{A}_2 = \{\xi^\top \Sigma \xi \geq r\lambda^2/2n\},$$

$$\mathbb{A}_3 = \{\langle \theta, \tilde{D}\xi\rangle^2 \leq 2\Delta^4\lambda^2\}.$$

Hence and since $\Delta^4 \geq \Delta^2$ *we get*

$$R(\eta^{**}) \geq Ce^{-c'n\Delta^4/r}(1 - \mathbf{P}(\mathbb{A}_1^c) - \mathbf{P}(\mathbb{A}_2^c) - \mathbf{P}(\mathbb{A}_3^c)).$$

Using hanson-Wright inequality it comes out that

$$\mathbf{P}(\mathbb{A}_1^c) + \mathbf{P}(\mathbb{A}_2^c) \leq 2e^{-cr} \leq 1/4,$$

since $r/n \geq 10$. *Moreover we also have that*

$$\mathbf{P}(\mathbb{A}_3^c) \leq \pi^D(e^{-c\Delta^4\lambda/\|\theta\|^2}) \leq e^{-c\Delta^2} + \mathbf{P}(\mathbb{A}_1^c) \leq 1/4,$$

since $\Delta^2 \geq 100$. *The proof is now complete.*

A.2. Proof of Theorem 2.2. *Let* θ *be a vector in* $\mathbb{R}^p$ *such that* $\|\theta\|^2 \geq \Delta^2\|\Sigma\|_\infty$. *Without loss of generality we may assume that* $\Delta^2 \geq C_1$ *for some constant* $C_1 > 0$ *large enough, otherwise the result is trivial as the upper bound becomes of constant order. We start by observing that*

$$\mathbb{P}((\hat{\eta}(Y_{n+1}) \neq \eta_{n+1}) = \mathbb{P}\left(\left\langle \sum_{i=1}^n Y_i\eta_i, Y_{n+1}\eta_{n+1}\right\rangle < 0\right).$$

Using the symmetry of the normal distribution (valid also with sub-Gaussian noise), we can assume that $\eta_{n+1} = 1$. *Let us denote by* $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n Y_i\eta_i = \theta + \xi$ *where* $\xi = \frac{1}{n}\sum_{i=1}^n W_i\eta_i$.*Then we get following upper bound*

$$\mathbb{P}((\hat{\eta}(Y_{n+1}) \neq \eta_{n+1}) = \mathbb{P}(\langle \hat{\theta}, \theta + W_{n+1}\rangle \leq 0) \leq C\mathbf{E}\left(e^{-c\frac{\langle \theta, \hat{\theta}\rangle^2}{\hat{\theta}^\top \Sigma \hat{\theta}}}\right).$$

for some constant $c, C > 0$ *where we have conditioned on* $\hat{\theta}$. *Next we have that*

$$\langle \hat{\theta}, \theta\rangle^2 = \langle \theta + \xi, \theta\rangle^2 \geq \frac{\|\theta\|^4}{2} - \langle \xi, \theta\rangle^2,$$

and that

$$\hat{\theta}^\top \Sigma \hat{\theta} \leq 2\|\Sigma\|_\infty \|\theta\|^2 + 2\xi^\top \Sigma \xi.$$

Hence

$$R(\hat{\eta}) \leq C\mathbf{E}\left(e^{-c\frac{\|\theta\|^4-2\langle\xi,\theta\rangle^2}{4\|\Sigma\|_\infty\|\theta\|^2+4\xi^\top\Sigma\xi}}\right).$$

Let us define now the random event

$$A = \{\xi^T\Sigma\xi \leq \mathrm{Tr}(\Sigma^2)/n + \|\Sigma\|_\infty\|\theta\|^2\} \cap \{4\langle\xi,\theta\rangle^2 \leq \|\theta\|^4\}.$$

Since $n\xi^T\Sigma\xi =_d z^T\Sigma^2 z = \|\Sigma z\|^2, \mathbb{E}\xi^T\Sigma\xi = \mathrm{Tr}(\Sigma^2)/n$ *and* $\sqrt{n}\xi^T\theta =_d \theta^T\Sigma^{1/2}z$ *where* $z \sim \mathcal{N}(0, I_p)$ *(or simply sub-Gaussian with i.i.d. entries). By Hanson-Wright inequality ([10]), for any* $t > 0$*,*

$$\mathbb{P}(n\xi^T\Sigma\xi - n\mathbb{E}\xi^T\Sigma\xi \geq t) = \mathbb{P}(z^T\Sigma^2 z - \mathrm{Tr}(\Sigma^2) \geq t) \leq e^{-c\min\{t^2/\mathrm{Tr}(\Sigma^4), t/\|\Sigma^2\|_\infty\}}$$

for some c*. Hence*

$$\mathbb{P}(A^c) \leq e^{-cn^2\|\Sigma\|_\infty^2\|\theta\|^4/\mathrm{Tr}(\Sigma^4)} + e^{-cn\|\theta\|^2/\|\Sigma\|_\infty} + e^{-cn\|\theta\|^4/\theta^\top\Sigma\theta},$$

for some $c > 0$ *small enough. Observing that* $\mathrm{Tr}(\Sigma^4) \leq \|\Sigma\|_\infty^2 \mathrm{Tr}(\Sigma^2)$ *we get further that*

$$\mathbb{P}(A^c) \leq 3e^{-c\frac{\Delta^4}{\Delta^2+\frac{r(\Sigma^2)}{n}}}.$$

Therefore we have

$$R(\hat{\eta}) \leq C\mathbf{E}\left(e^{-c\frac{\|\theta\|^4-2\langle\xi,\theta\rangle^2}{4\|\Sigma\|_\infty\|\theta\|^2+4\xi^\top\Sigma\xi}}\mathbf{1}\{A\}\right) + 3Ce^{-c\frac{\Delta^4}{\Delta^2+\frac{r(\Sigma^2)}{n}}}.$$

We conclude, using the event A*, that*

$$R(\hat{\eta}) \leq C\exp\left(-c\frac{\Delta^4}{\Delta^2+\frac{r(\Sigma^2)}{n}}\right)$$

for some $c, C > 0$*.*

APPENDIX B: PROOFS OF REGULARIZATION VS INTERPOLATION

B.1. Proof of Theorem 3.1. *Recall that* W_{n+1} *is* $\mathcal{N}(0,\Sigma)$ *(or sub-Gaussian). Hence for any vector* v *the random variable* $v^\top W_{n+1}$ *is a centered Gaussian of variance* $v^\top\Sigma v$*. It follows that*

$$\mathbb{P}\left(\hat{\eta}(Y_{n+1}) \neq \eta_{n+1}\right) \leq Ce^{-\frac{\langle\theta,\hat{\theta}_\lambda\rangle^2}{2\hat{\theta}_\lambda^\top\Sigma\hat{\theta}_\lambda}},$$

conditionally on $\hat{\theta}_\lambda$. Observe that $\hat{\theta}_\lambda = \theta x/n + W^\top A_\lambda^{-1}\eta/n$ where $A_\lambda = \lambda I_n + Y^T Y/n$ and $x = \eta^T A_\lambda^{-1}\eta$. The risk is invariant by rescaling $\hat{\theta}_\lambda$ hence we rescale it by n/x. Hence without loss of generality we may assume that $\hat{\theta}_\lambda = \theta + W^\top H_\lambda^{-1}\eta/x$. Using Lemma 3 we have

$$\hat{\theta}_\lambda = \left(I_p - WA^{-1}W^\top/n\right)\theta + \frac{1+\eta^\top A^{-1}W^\top\theta/n}{\eta^\top A^{-1}\eta} WA^{-1}\eta.$$

On the one hand we have

$$|\langle\theta, \hat{\theta}_\lambda\rangle| \geq \|\theta\|^2 - \theta^\top WA^{-1}W^\top\theta/n - \frac{|\eta^\top A^{-1}W^\top\theta|}{\eta^\top A^{-1}\eta}.$$

On the other hand we have

$$\hat{\theta}_\lambda^\top \Sigma\hat{\theta}_\lambda \leq 2\left(\|\Sigma^{1/2}\left(I_p - WA^{-1}W^\top/n\right)\theta\|^2 + \frac{2+2(\eta^\top A^{-1}W^\top\theta/n)^2}{(\eta^\top A^{-1}\eta)^2}\eta^T A^{-1}W^T\Sigma WA^{-1}\eta\right).$$

- *Control of the numerator:*

$$|\langle\theta, \hat{\theta}_\lambda\rangle| \geq \|\theta\|^2 - \|A^{-1/2}W^\top\theta\|^2/n - \frac{\|A^{-1/2}W^\top\theta\|}{\sqrt{\eta^\top A^{-1}\eta}}.$$

Using Lemma 6 and Lemma 7 we get that

$$|\langle\theta, \hat{\theta}_\lambda\rangle| \geq \|\theta\|^2 - C_1\frac{\theta^\top\Sigma\theta}{\sum_{i>k}\lambda_i/n+\lambda} - C_2\|\theta\|_\Sigma.$$

- *Control of the denominator:*

$$\hat{\theta}_\lambda^\top \Sigma\hat{\theta}_\lambda \leq C\left(\|\theta\|_\Sigma^2 + \|\Sigma^{1/2}WA^{-1}W^\top\theta\|^2/n^2 + \frac{2+2(\eta^\top A^{-1}W^\top\theta/n)^2}{(\eta^\top A^{-1}\eta)^2}\eta^T A^{-1}W^T\Sigma WA^{-1}\eta\right).$$

Using the same bound as for the numerator and the fact that $\|W^\top\theta\|^2 \leq Cn\|\theta\|_\Sigma^2$ with probability $1-e^{-cn}$ we get further that

$$\hat{\theta}_\lambda^\top \Sigma\hat{\theta}_\lambda \leq C\left(\|\theta\|_\Sigma^2(1+\|\Sigma^{1/2}WA^{-1}\|_\infty^2/n) + ((\sum_{i>k}\lambda_i/n+\lambda)^2 + \|\theta\|_\Sigma^2)\frac{\eta^T A^{-1}W^T\Sigma WA^{-1}\eta}{n^2}\right).$$

Using Lemma 5 we have that with probability $1-\delta$

$$\eta^T A^{-1}W^T\Sigma WA^{-1}\eta \leq 3/2\,\mathrm{Tr}(A^{-1}W^T\Sigma WA^{-1}) + \|A^{-1}W^T\Sigma WA^{-1}\|_\infty \log(1/\delta).$$

Hence using Lemma 8 and Lemma 9 we get that with probability $1-\delta$

$$\eta^T A^{-1}W^T\Sigma WA^{-1}\eta \leq C\left(k^*n + n\frac{\sum_{i>k}\lambda_i^2}{(\sum_{i>k}\lambda_i/n+\lambda)^2}\right) + \left(k^*n + \frac{\sum_{i>k}\lambda_i^2 + \lambda_{k+1}^2 n}{(\sum_{i>k}\lambda_i/n+\lambda)^2}\right)\log(1/\delta),$$

and that

$$\|\Sigma^{1/2}WA^{-1}\|_\infty^2/n \leq C(k^* + \frac{\sum_{i>k}\lambda_i^2/n + \lambda_{k+1}^2}{(\sum_{i>k}\lambda_i/n+\lambda)^2}) \leq C(1+k^*).$$

Notice also that

$$\frac{\eta^T A^{-1}W^T\Sigma W A^{-1}\eta}{n^2} \leq C_1 + C_2\,(k^*+1)\,/n\log(1/\delta).$$

We conclude that with probability $1-\delta_1-\delta_2$

$$\hat\theta_\lambda^\top\Sigma\hat\theta_\lambda \leq C\left(\|\theta\|_\Sigma^2\,(1+k^*+(1+k^*)/n\log(1/\delta_1)) + \frac{k^*\lambda_k^2+\sum_{i>k}\lambda_i^2}{n} + \frac{(k^*\lambda_k^2+\lambda_{k+1}^2)\log(1/\delta_2)}{n}\right)$$

Hence by taking $\delta_1 = e^{-cn}$ *we get further that*

$$\hat\theta_\lambda^\top\Sigma\hat\theta_\lambda \leq C\left(\|\theta\|_\Sigma^2(1+k^*) + \frac{k^*\lambda_k^2+\sum_{i>k}\lambda_i^2}{n} + \frac{(k^*\lambda_k^2+\lambda_{k+1}^2)\log(1/\delta_2)}{n}\right).$$

We have used the fact that k *the smallest integer that satisfies* $r_k(\Sigma)+\lambda/\lambda_{k+1} > bn$ *for* $b \geq 1$. *Hence we have*

$$\sum_{i\geq k}\lambda_i/n+\lambda \leq b\lambda_k.$$

Notice that k *is a decreasing function of* λ. *We treat two cases. If* $\mathrm{Tr}(\Sigma)/n+\lambda \geq \|\Sigma\|_\infty$ *then we can take* $k=0$. *Now if* $\mathrm{Tr}(\Sigma)/n+\lambda \leq \|\Sigma\|_\infty$, *then* $k \geq 1$ *and we use the above property of* k

$$\frac{k(\sum_{i>k}\lambda_i/n+\lambda)^2}{n^3} + n^2\frac{\sum_{i>k}\lambda_i^2}{n^3} \leq \frac{bkn^2\lambda_k^2}{n^3} + \frac{\sum_{i>k}\lambda_i^2}{n} \leq C\frac{k\lambda_k^2+\sum_{i>k}\lambda_i^2}{n}.$$

Since k *is a decreasing function of* λ *then* $k\lambda_k^2+\sum_{i>k}\lambda_i^2$ *is also a decreasing function of* λ. *Observe that*

$$k\lambda_k^2 + \sum_{i>k}\lambda_i^2 \leq \mathrm{Tr}\,(\Sigma^2).$$

B.2. Proof of Theorem 4.1. *Using the proof of Lemma 3, we have*

$$n(Y^\top Y)^{-1}\eta = \frac{\sqrt{n}}{\|\theta\| det}\left(A^{-1}u(1+u^\top A^{-1}v) - A^{-1}vu^\top A^{-1}u\right),$$

where $A = W^\top W/n$, $u = \|\theta\|\eta/\sqrt{n}$ and $v = W^\top\theta/\sqrt{n\|\theta\|^2}$. Hence $e_1^\top(Y^\top Y)^{-1}\eta$ has the sign as

$$e_1^\top(W^\top W)^{-1}\eta(1+\eta^\top(W^\top W)^{-1}W^\top\theta) - e_1^\top(W^\top W)^{-1}W^\top\theta\eta^\top(W^\top W)^{-1}\eta.$$

Using Lemma 4 we also have

$$e_1(W^\top W)^{-1}\omega = \frac{\omega_1 - W_1^\top\tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\omega}}{\|W_1\|^2 - W_1^\top\pi W_1}.$$

Notice that $\pi = \tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{W}^\top$ is a $p\times p$ projection matrix. Since $W \sim N(0,\Sigma)$, W_1 is independent of the column space spanned by $\tilde{W}$. Therefore $\|W_1\|^2 - W_1^\top\pi W_1 = W_1^\top(I_p-\pi)W_1$ is always positive.

Hence we only need to show that the following expression is positive

$$(1-\eta_1 W_1^\top\tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\eta})(1+\eta^\top(W^\top W)^{-1}W^\top\theta) - \eta_1 W_1^\top(I_p-\pi)\theta\eta^\top(W^\top W)^{-1}\eta.$$

We first use the bound

$$\eta^\top(W^\top W)^{-1}W^\top\theta \le \sqrt{\eta A^{-1}\eta}\|A^{-1/2}W^\top\theta\|/n \le C\frac{\sqrt{\theta^\top\Sigma\theta}}{\sum_{i>k}\lambda_i/n}.$$

Hence under the condition $\|\theta\|_\Sigma \le C\sum_{i>k}\lambda_i/n$ we have that

$$1+\eta^\top(W^\top W)^{-1}W^\top\theta \ge 1/2.$$

Next observe that $\eta_1 W_1^\top(I_p-\pi)\theta$ is sub-Gaussian with parameter $\|\Sigma^{1/2}(I_p-\pi)\theta\|$. Hence using the bound from the proof of Theorem 3.1, we get that

$$\|\Sigma^{1/2}(I_p-\pi)\theta\|^2 \le C\|\theta^\top\Sigma\theta\|^2(1+k^*).$$

It comes out that with probability at least $1-\delta$ we have

$$|\eta_1 W_1^\top(I_p-\pi)\theta\eta^\top(W^\top W)^{-1}\eta| \le C\frac{\|\theta\|_\Sigma\sqrt{(1+k^*)\log(1/\delta)}}{\sum_{i>k}\lambda_i/n} \le 1/4,$$

under our condition $\|\theta\|_\Sigma\sqrt{(1+k^*)\log(1/\delta)} \le C\sum_{i>k}\lambda_i/n$. Finally, we have that $\eta_1 W_1^\top\tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\eta}$ is sub-Gaussian with parameter $\|\Sigma^{1/2}\tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\eta}\|$. Using the proof of Theorem 3.1 we have

$$\|\Sigma^{1/2}\tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\eta}\|^2 \le C\left(k^*/n + \frac{\sum_{i>k}\lambda_i^2}{n(\sum_{i>k}\lambda_i/n)^2}\right) + \left(k^*/n + \frac{\sum_{i>k}\lambda_i^2/n^2 + \lambda_{k+1}^2/n}{(\sum_{i>k}\lambda_i/n)^2}\right)\log(1/\delta).$$

Hence

$$\|\Sigma^{1/2}\tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\eta}\|^2 \leq C\left(k^*\log(1/\delta)/n + \frac{(\sum_{i>k}\lambda_i^2 + \lambda_{k+1}^2\log(1/\delta))/n}{(\sum_{i>k}\lambda_i/n)^2}\right).$$

Hence with probability $1 - C/n^2$ *we have*

$$|\eta_1 W_1^\top\tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\eta}| \leq C\left(\sqrt{k^*\log^2(n)/n} + \frac{\sqrt{\sum_{i>k}\lambda_i^2\log(n)/n} + \lambda_{k+1}\log(n)/\sqrt{n}}{\sum_{i>k}\lambda_i/n}\right).$$

Hence under the conditions $k^*\log^2(n) \leq Cn$ *and* $\sum_{i>k}\lambda_i^2 n\log(n) \leq C(\sum_{i>k}\lambda_i)^2$ *we get our result. We can now Conclude using the union bound for* $i = 1,\dots,n$*. Hence the final bound hold with probability* $1 - 1/n$*.*

APPENDIX C: AUXILIARY RESULTS (ALGEBRA)

LEMMA 1. *Let* $u, v \in \mathbb{R}^n$ *and* $A \in \mathbb{R}^{n\times n}$ *an invertible and symmetric matrix then*

$$(uu^\top + uv^\top + vu^\top + A)^{-1} - A^{-1} = -\frac{(1 - v^\top A^{-1}v)A^{-1}uu^\top A^{-1} + (1 + u^\top A^{-1}v)A^{-1}(uv^\top + vu^\top)A^{-1} - u^\top A^{-1}uA^{-1}vv^\top A^{-1}}{(1 - v^\top A^{-1}v)u^\top A^{-1}u + (1 + u^\top A^{-1}v)^2}.$$

PROOF. We will use the Woodbury matrix identity

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1},$$

with $U = (u\ v) \in \mathbb{R}^{n\times 2}$, $V = U^\top$ and $C = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$. We start by computing $C^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}$. It comes out that

$$C^{-1} + VA^{-1}U = \begin{pmatrix} u^\top A^{-1}u & 1 + u^\top A^{-1}v \\ 1 + u^\top A^{-1}v & v^\top A^{-1}v - 1 \end{pmatrix}.$$

Let us denote by $det = (1 - v^\top A^{-1}v)u^\top A^{-1}u + (1 + u^\top A^{-1}v)^2$, then

$$(C^{-1} + VA^{-1}U)^{-1} = \frac{1}{det}\begin{pmatrix} 1 - v^\top A^{-1}v & 1 + u^\top A^{-1}v \\ 1 + u^\top A^{-1}v & -u^\top A^{-1}u \end{pmatrix}.$$

Moreover

$$U(C^{-1}+VA^{-1}U)^{-1}V = \frac{(1 - v^\top A^{-1}v)uu^\top + (1 + u^\top A^{-1}v)(uv^\top + vu^\top) - u^\top A^{-1}uvv^\top}{det}.$$

We conclude observing that

$$uu^\top + uv^\top + vu^\top + A = A + UCV.$$

□

LEMMA 2. *Let $u, v \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ and an invertible and symmetric matrix then*

$$(uu^\top + uv^\top + vu^\top + A)^{-1}u = \frac{A^{-1}u(1 + u^\top A^{-1}v) - A^{-1}vu^\top A^{-1}u}{(1 - v^\top A^{-1}v)u^\top A^{-1}u + (1 + u^\top A^{-1}v)^2},$$

and

$$(uu^\top + uv^\top + vu^\top + A)^{-1}v = \frac{A^{-1}v(1 + u^\top A^{-1}v + u^\top A^{-1}u) - A^{-1}u(u^\top A^{-1}v + v^\top A^{-1}v)}{(1 - v^\top A^{-1}v)u^\top A^{-1}u + (1 + u^\top A^{-1}v)^2}.$$

PROOF. Using Lemma 1. □

LEMMA 3. *For any $\lambda > 0$, we have that*

$$\hat{\theta}_\lambda = \left(I_n - WA^{-1}W^\top/n\right)\theta + \frac{1 + \eta^\top A^{-1}W^\top\theta/n}{\eta^\top A^{-1}\eta}WA^{-1}\eta,$$

where $A = \lambda I_n + W^\top W/n$. We also have

$$\hat{\theta}_\lambda/c = (1 - \eta^\top W^\top B^{-1}W\eta/n^2)B^{-1}\theta + (1 + \theta^\top B^{-1}W\eta/n)B^{-1}W\eta/n,$$

where $B = \lambda I_p + WW^\top/n$ and $c > 0$ some constant.

PROOF. Recall that $\hat{\theta}_\lambda = \theta + WH_\lambda^{-1}\eta/x$ where $A_\lambda = \lambda I_n + Y^\top Y/n$ and $x = \eta^T A_\lambda^{-1}\eta$. By denoting $w = \theta/\|\theta\|$, we have that

$$H_\lambda = \frac{\|\theta\|^2}{n}\eta\eta^\top + \frac{\|\theta\|}{n}(\eta(W^\top w)^\top + (W^\top w)\eta^\top) + \lambda I_n + W^\top W/n.$$

By choosing $u = \|\theta\|\eta/\sqrt{n}$, $v = W^\top w/\sqrt{n}$ and $A = \lambda I_n + W^\top W/n$ we get using Lemma 2 that

$$H_\lambda^{-1}\eta = \frac{\sqrt{n}}{\|\theta\| det}\left(A^{-1}u(1 + u^\top A^{-1}v) - A^{-1}vu^\top A^{-1}u\right),$$

where $det = (1 - v^\top A^{-1}v)u^\top A^{-1}u + (1 + u^\top A^{-1}v)^2$. It comes out that

$$\eta^\top H_\lambda^{-1}\eta = \frac{nu^\top A^{-1}u}{\|\theta\|^2 det} = \frac{\eta^\top A^{-1}\eta}{det}.$$

We also get

$$\begin{aligned} WH_\lambda^{-1}\eta &= \frac{\sqrt{n}}{\|\theta\| det}\left(WA^{-1}u(1+u^\top A^{-1}v) - WA^{-1}vu^\top A^{-1}u\right) \\ &= \frac{1}{det}\left(WA^{-1}\eta(1+\eta^\top A^{-1}W^\top\theta/n) - \frac{\eta^\top A^{-1}\eta}{n}WA^{-1}W^\top\theta\right). \end{aligned}$$

As a conclusion we get that

$$\hat{\theta}_\lambda det = \left(I_n - WA^{-1}W^\top/n\right)\theta + \frac{1+\eta^\top A^{-1}W^\top\theta/n}{\eta^\top A^{-1}\eta}WA^{-1}\eta.$$

On the other hand we also have

$$\hat{\theta}_\lambda = \left(\theta\theta^\top + \theta(W\eta)^\top/n + W\eta\theta^\top/n + \lambda I_p + WW^\top/n\right)^{-1}(\theta + W\eta/n).$$

By choosing $u = \theta$, $v = W\eta/n$ and $B = \lambda I_p + WW^\top/n$ we get using Lemma 2 that

$$\begin{aligned} \hat{\theta}_\lambda &= (uu^\top + uv^\top + vu^\top + B)^{-1}(u+v) \\ &= \frac{1}{det}\left(B^{-1}u(1-v^\top B^{-1}v) + B^{-1}v(1+u^\top A^{-1}v)\right). \end{aligned}$$

where $det = (1-v^\top B^{-1}v)u^\top B^{-1}u + (1+u^\top B^{-1}v)^2$. Hence

$$\hat{\theta}_\lambda det = (1-\eta^\top W^\top B^{-1}W\eta/n^2)B^{-1}\theta + (1+\theta^\top B^{-1}W\eta/n)B^{-1}W\eta/n.$$

□

Lemma 4. *Let $W = (W_1\ \tilde{W})$ and $\omega = (\omega_1\ \tilde{\omega})$, then we have*

$$e_1(W^\top W)^{-1}\omega = \frac{\omega_1 - W_1^\top\tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\omega}}{\|W_1\|^2 - W_1^\top\pi W_1},$$

where $\pi = \tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{W}^\top$.

Proof. We will use the following formula (Schur complement) that holds as long as all matrix inverses exist.

$$\begin{pmatrix} A & B \\ B^\top & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D-B^\top A^{-1}B)^{-1}B^\top A^{-1} & -A^{-1}B(D-B^\top A^{-1}B)^{-1} \\ -(D-B^\top A^{-1}B)^{-1}B^\top A^{-1} & (D-B^\top A^{-1}B)^{-1} \end{pmatrix}.$$

Considering $A = \|W_1\|^2, B = W_1^\top\tilde{W}$ and $D = \tilde{W}^\top\tilde{W}$, then

$$W^\top W = \begin{pmatrix} A & B \\ B^\top & D \end{pmatrix}$$

By Sherman-Morrison formula we have that

$$\begin{aligned}(D - B^\top A^{-1}B)^{-1} &= \left(\tilde{W}^\top \tilde{W} - \frac{1}{\|W_1\|^2}\tilde{W}^\top W_1 W_1^\top \tilde{W}\right)^{-1} \\ &= (\tilde{W}^\top \tilde{W})^{-1} + \frac{\frac{1}{\|W_1\|^2}(\tilde{W}^\top \tilde{W})^{-1}\tilde{W}^\top W_1 W_1^\top \tilde{W}(\tilde{W}^\top \tilde{W})^{-1}}{1 - \frac{W_1^\top \pi W_1}{\|W_1\|^2}} \\ &= (\tilde{W}^\top \tilde{W})^{-1} + \frac{(\tilde{W}^\top \tilde{W})^{-1}\tilde{W}^\top W_1 W_1^\top \tilde{W}(\tilde{W}^\top \tilde{W})^{-1}}{\|W_1\|^2 - W_1^\top \pi W_1} \\ &= \frac{E}{\|W_1\|^2 - W_1^\top \pi W_1}.\end{aligned}$$

where $E = (\tilde{W}^\top \tilde{W})^{-1}(\|W_1\|^2 - W_1^\top \pi W_1) + (\tilde{W}^\top \tilde{W})^{-1}\tilde{W}^\top W_1 W_1^\top \tilde{W}(\tilde{W}^\top \tilde{W})^{-1}$. Hence

$$B(D - B^\top A^{-1}B)^{-1} = \frac{W_1^\top \tilde{W}(\tilde{W}^\top \tilde{W})^{-1}}{1 - \frac{W_1^\top \pi W_1}{\|W_1\|^2}} = \frac{\|W_1\|^2 W_1^\top \tilde{W}(\tilde{W}^\top \tilde{W})^{-1}}{\|W_1\|^2 - W_1^\top \pi W_1}.$$

It comes out that

$$A^{-1} + A^{-1}B(D - B^\top A^{-1}B)^{-1}B^\top A^{-1} = \frac{1}{\|W_1\|^2}\frac{1}{1 - \frac{W_1^\top \pi W_1}{\|W_1\|^2}} = \frac{1}{\|W_1\|^2 - W_1^\top \pi W_1},$$

and that

$$A^{-1}B(D - B^\top A^{-1}B)^{-1} = \frac{1}{\|W_1\|^2}\frac{W_1^\top \tilde{W}(\tilde{W}^\top \tilde{W})^{-1}}{1 - \frac{W_1^\top \pi W_1}{\|W_1\|^2}} = \frac{W_1^\top \tilde{W}(\tilde{W}^\top \tilde{W})^{-1}}{\|W_1\|^2 - W_1^\top \pi W_1}.$$

Hence

$$(W^\top W)^{-1} = \frac{1}{\|W_1\|^2 - W_1^\top \pi W_1}\begin{pmatrix} 1 & -W_1^\top \tilde{W}(\tilde{W}^\top \tilde{W})^{-1} \\ -(\tilde{W}^\top \tilde{W})^{-1}\tilde{W}^\top W_1 & E \end{pmatrix}$$

and

$$e_1(W^\top W)^{-1}\omega = \frac{\omega_1 - W_1^\top \tilde{W}(\tilde{W}^\top \tilde{W})^{-1}\tilde{\omega}}{\|W_1\|^2 - W_1^\top \pi W_1}.$$

□

APPENDIX D: AUXILIARY RESULTS (PROBABILITY)

Following the idea of [2], we write Σ *in its eigen-decomposition form* $\Sigma = \sum_{i=1}^{n} \lambda_i v_i v_i^T$ *and decompose* $W^\top W$ *in the basis of* Σ*. More specifically, let* $z_i = W^T v_i / \sqrt{\lambda_i}$*. Then* $\{z_i\}$ *is a basis (not orthogonal, if we assume* W *full rank) of* $\mathbb{R}^{n\times n}$ *and* $W^T W = \sum_i \lambda_i z_i z_i^T$*. With our notations* $A_\lambda = \frac{1}{n} W^\top W + \lambda I_n = \sum_i \lambda'_i z_i z_i^T + \lambda I_n$ *where* $\lambda'_i = \lambda/n$*. Let us also define*

$$A_{-i} = \sum_{j\neq i} \lambda'_j z_j z_j^T + \lambda I_n, \quad A_k = \sum_{i>k} \lambda'_i z_i z_i^T + \lambda I_n,$$

and similarly

$$A_{-i}^0 = A_{-i} - \lambda I_n, \quad A_k^0 = A_k - \lambda I_n.$$

In that case we can apply Lemma 9 and 10 still to A_{-i}^0 *and* A_k^0*. Adding* λ *to all sides, we get with probability at least* $1 - 2e^{-n/c}$ *that:*

- *For any* $k \geq 0$

$$\frac{1}{c}\sum_{i>k} \lambda'_i - c\lambda'_{k+1} n + \lambda \leq \theta_n(A_k) \leq \theta_1(A_k) \leq c\left(\sum_{i>k} \lambda'_i + \lambda'_{k+1} n\right) + \lambda$$

- $\forall i \geq 1$,

$$\theta_{k+1}(A_{-i}) \leq \theta_{k+1}(A) \leq \theta_1(A_k) \leq c\left(\sum_{i>k} \lambda'_i + \lambda'_{k+1} n\right) + \lambda$$

- $1 \leq i \leq k$,

$$\theta_n(A) \geq \theta_n(A_{-i}) \geq \theta_n(A_k) \geq \frac{1}{c}\sum_{i>k} \lambda'_i - c\lambda'_{k+1} n + \lambda$$

Recall that the effective rank is given by $r_k(\Sigma) := \frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}}$*. Hence under the condition* $r_k(\Sigma) + \frac{n\lambda}{\lambda_{k+1}} \geq bn$ *the above inequalities become*

- *for* $k = k^*$

$$\frac{1}{c}\sum_{i>k} \lambda_i/n + \lambda \leq \theta_n(A_k) \leq \theta_1(A_k) \leq c\sum_{i>k} \lambda_i/n + \lambda$$

- $\forall i \geq 1$,

$$\theta_{k+1}(A_{-i}) \leq \theta_{k+1}(A) \leq \theta_1(A_k) \leq c\sum_{i>k} \lambda_i/n + \lambda$$

- $1 \leq i \leq k$,

$$\theta_n(A_\lambda) \geq \theta_n(A_{-i}) \geq \theta_n(A_k) \geq \frac{1}{c} \sum_{i>k} \lambda_i/n + \lambda$$

LEMMA 5. *Let ξ be a 1 sub-Gaussian vector with i.i.d entries and let B be a SDP matrix. Then with probability at least $1 - \delta$, we have*

$$|\xi^\top B \xi - \mathrm{Tr}(B)| \leq 1/2\, \mathrm{Tr}(B) + C\|B\|_\infty \log(1/\delta),$$

for some $C > 0$.

PROOF. Using Hanson-Wright inequality. □

LEMMA 6. *With probability at least $1 - e^{-cn}$ we have*

$$\|A^{-1/2} W^\top \theta\|^2 \leq C \frac{n\theta^\top \Sigma \theta}{\sum_{i>k} \lambda_i/n + \lambda}.$$

PROOF. We have that

$$\|A^{-1/2} W^\top \theta\|^2 \leq \|W^\top \theta\|_2^2 \|A^{-1}\|_\infty.$$

Since $W^\top \theta$ has the same distribution as $\|\theta\|_\Sigma.\xi$ where ξ is 1 sub-Gaussian with i.i.d entries, then using Lemma 5 we get with probability at least $1 - e^{-n}$

$$\|A^{-1/2} W^\top \theta\|^2 \leq C\theta^\top \Sigma \theta n \|A^{-1}\|_\infty.$$

Hence we conclude that

$$\|A^{-1/2} W^\top \theta\|^2 \leq C \frac{n\theta^\top \Sigma \theta}{\sum_{i>k} \lambda_i/n + \lambda}.$$

□

LEMMA 7. *There exist $C_1, C_2 > 0$ such that*

$$\eta^\top A^{-1} \eta \leq C_1 \frac{n}{\sum_{i>k} \lambda_i/n + \lambda},$$

and with probability at least $1 - e^{-cn}$ we have

$$\eta^\top A^{-1} \eta \geq C_2 \frac{n}{\sum_{i>k} \lambda_i/n + \lambda}.$$

Proof. The first inequality is straightforward observing that

$$\eta^\top A^{-1}\eta \le n\|A^{-1}\|_\infty.$$

For the lower bound, we will the sub-Gaussian property of η. To lower bound $\mathrm{Tr}(A^{-1})$ observe that with probability $1-e^{-cn}$ we have

$$\mathrm{Tr}(A^{-1}) = \sum_{i=1}^{n}(\theta_i(A))^{-1} \ge \sum_{i=k+1}^{n}\frac{1}{c\lambda_{k+1}r_k(\Sigma)/n+\lambda} \ge \frac{cn}{\lambda_{k+1}r_k(\Sigma)/n+\lambda}.$$

We conclude using Lemma 5. □

Lemma 8. *With probability at least* $1-e^{-cn}$ *we have*

$$\mathrm{Tr}(A^{-1}W^T\Sigma W A^{-1}) \le c\left(k^*n + n\frac{\sum_{i>k}\lambda_i^2}{(\sum_{i>k}\lambda_i/n+\lambda)^2}\right).$$

Proof. Let us denote by $C := A^{-1}W^T\Sigma W A^{-1}$. Using the rank one inverse formula we get

$$\begin{aligned}\mathrm{Tr}(C) &= \mathrm{Tr}(A^{-1}W^T\Sigma W A^{-1})\\ &= \sum_{i=1}^{n}\lambda_i^2 z_i(\lambda_i' z_i z_i^T + A_{-i})^{-2}z_i^T\\ &= \sum_{i=1}^{n}\frac{\lambda_i^2 z_i^T A_{-i}^{-2}z_i}{(1+\lambda_i' z_i^T A_{-i}^{-1}z_i)^2}.\end{aligned}$$

Then for some $l \le k^*$,

$$\mathrm{Tr}(C) = \sum_{i=1}^{l}\frac{\lambda_i^2 z_i^T A_{-i}^{-2}z_i}{(1+\lambda_i' z_i^T A_{-i}^{-1}z_i)^2} + \sum_{i>l}\lambda_i^2 z_i A_\lambda^{-2}z_i^T$$

Under the condition $r_k(\Sigma)+n\lambda/\lambda_{k+1} \ge bn$, there exists c_1 such that with probability at least $1-2e^{-n/c_1}$, for $i \le k$, $\theta_n(A_{-i}) \ge \frac{1}{c}\sum_{i>k}\lambda_i/n+\lambda$. Hence

$$z^T A_{-i}^{-2}z \le \frac{c_1^2\|z\|^2}{\left(\sum_{i>k}\lambda_i/n+\lambda/c_1\right)^2}.$$

Let $\mathscr{L}_i$ be the span of eigenvectors of A corresponding to $n-k^*$ smallest eigenvalues. Then

$$z^T A_{-i}^{-1}z_i \ge \left(\Pi_{\mathscr{L}_i}z\right)^T A_{-i}^{-1}\Pi_{\mathscr{L}_i}z \ge \frac{\left\|\Pi_{\mathscr{L}_i}z\right\|^2}{1/c\sum_{i>k}\lambda_i/n+\lambda}$$

Using the sub-Gaussian property, it comes out that with probability $1 - 3e^{-cn}$, for all i

$$\|z_i\|^2 \leq n$$
$$\|\Pi_{\mathscr{L}_i} z_i\|^2 \geq n.$$

Hence the first sum can be bounded by

$$\sum_{i=1}^{l} \frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1 + \lambda_i' z_i^T A_{-i}^{-1} z_i)^2} \leq c_1^4 \frac{n^2 \sum_{i=1}^{l} \|z_i\|^2}{\|\Pi_{\mathscr{L}_i} z_i\|^4} \leq c_4 l n.$$

For the second sum, consider the same event we have $\theta_n(A) \geq \lambda_{k+1} r_k(\Sigma)/nc_1 + \lambda$,

$$\sum_{i>l} \lambda_i^2 z_i^T A^{-2} z_i \leq \frac{c_1^2 \sum_{i>l} \lambda_i^2 \|z_i\|^2}{(\sum_{i>k} \lambda_i/n + \lambda/c_1)^2} \leq \frac{c_5 n \sum_{i>l} \lambda_i^2}{(\sum_{i>k} \lambda_i/n + \lambda/c_1)^2}$$

Therefore we have

$$\mathrm{Tr}(C) \leq c \left(k^* n + n \frac{\sum_{i>k} \lambda_i^2}{(\sum_{i>k} \lambda_i/n + \lambda)^2} \right)$$

for $0 \leq k \leq n/c$ with high probability.

□

Lemma 9. *With probability at least* $1 - e^{-cn}$ *we have*

$$\|A^{-1} W^T \Sigma W A^{-1}\|_\infty \leq c \left(k^* n + \frac{\sum_{i>k} \lambda_i^2 + \lambda_{k+1}^2 n}{(\sum_{i>k} \lambda_i/n + \lambda)^2} \right).$$

Proof. For the first half (elements $i \leq k$) we simply bound spectral norm with Trace. It only remains to bound

$$\|\sum_{i>k} \lambda_i^2 A^{-1} z_i z_i^\top A^{-1}\|_\infty \leq n^2 \|A^{-1}\|_\infty^2 \|A_k\|_\infty.$$

Hence we get

$$\|\sum_{i>k} \lambda_i^2 A^{-1} z_i z_i^\top A^{-1}\|_\infty \leq \frac{c\left(\sum_{i>k} \lambda_i^2 + \lambda_{k+1}^2 n\right)}{(\sum_{i>k} \lambda_i/n + \lambda)^2}.$$

We conclude that

$$\|A^{-1} W^T \Sigma W A^{-1}\|_\infty \leq c \left(k^* n + \frac{\sum_{i>k} \lambda_i^2 + \lambda_{k+1}^2 n}{(\sum_{i>k} \lambda_i/n + \lambda)^2} \right).$$

□

Department of Mathematics
University of Southern California
Los Angeles, CA 90089
E-mail: yiqiushe@usc.edu; minsker@usc.edu; ndaoud@usc.edu