Journal of Machine Learning Research 26 (2025) 1-39 Submitted 8/24; Revised 3/25; Published 7/25

Classification in the high dimensional Anisotropic mixture framework: A new take on Robust Interpolation

Stanislav Minsker MINSKER@USC.EDU
Department of Mathematics
University of Southern California
Los Angeles, CA 90089

Mohamed Ndaoud NDAOUD@ESSEC.EDU
Department of Information Systems, Decision Sciences and Statistics
ESSEC Business School
95000 Cergy, France

Yiqiu Shen YIQIUSHE@USC.EDU
Department of Data Sciences and Operations
University of Southern California
Los Angeles, CA 90089

Editor: Joan Bruna

Abstract

We study the classification problem under the two-component anisotropic sub-Gaussian mixture model in high dimensions and in the non-asymptotic setting. First, we derive lower bounds and matching upper bounds for the minimax risk of classification in this framework. We also show that in the high-dimensional regime, the linear discriminant analysis classifier turns out to be sub-optimal in the minimax sense. Next, we give precise characterization of the risk of classifiers based on solutions of ℓ_2-regularized least squares problem. We deduce that the interpolating solutions may outperform the regularized classifiers under mild assumptions on the covariance structure of the noise, and present concrete examples of this phenomenon. Our analysis also demonstrates robustness of interpolation to certain models of corruption. To the best of our knowledge, this peculiar fact 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.

Keywords: benign overfitting, minimax classification, high-dimensional statistics, robustness, regularization

1. Introduction

The topic of overparameterization has gained increasing attention in recent literature on problems in high-dimensional statistics. Previously, it was widely believed that estimators given by solutions to properly regularized optimization problems, discussed for instance in (Bickel et al., 2006; Wainwright, 2014; De Vito et al., 2021), yield the best generalization power. Recently, it was discovered (see Belkin et al., 2019b; Oymak and Soltanolkotabi, 2019; Liang and Rakhlin, 2020; Hastie et al., 2022, among other works) that the estimators

which interpolate the training data also yield good generalization error bounds when the number of covariates exceeds the sample size. This phenomenon, termed "benign overfitting" Bartlett et al. (2020); Chinot and Lerasle (2021), has been extensively investigated in the regression and classification settings Chinot et al. (2021); Liang and Recht (2021). A closely related notion of "double descent" which, broadly speaking, refers to the situation when both the models with relatively low complexity as well as the models with very high complexity that perfectly interpolate the training data achieve small generalization error Belkin et al. (2019a). In this work, we focus on the high-dimensional classification problem. We derive the bounds for the generalization error under different sets of assumptions and compare the interpolating solution to the estimators that have previously been studied in the literature. We emphasize the different phenomena that play key roles in the low dimensional and the high dimensional settings.

1.1 Notation and definitions

For $\theta \in \mathbb{R}^p$, we denote its Euclidean norm by $\|\theta\|$ and its Mahalanobis norm corresponding to a positive definite matrix Σ by $\|\theta\|_\Sigma^2 := \theta^\top \Sigma^{-1}\theta$. For a symmetric matrix $A \in \mathbb{R}^{p\times p}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$, we define the spectral norm of A via $\|A\|_\infty := \lambda_1$ and the Frobenius norm by $\|A\|_F := \sqrt{\sum_{i=1}^p \lambda_i^2}$. For a positive semidefinite (PSD) matrix $A = \sum_{i=1}^p \lambda_i v_i v_i^\top \neq 0_p, A \in \mathbb{R}^{p\times p}$, we define its trace via $\mathrm{Tr}(A) = \sum_{i=1}^p \lambda_i$, its effective rank by

$$r(A) := \frac{\mathrm{Tr}(A)}{\|A\|_\infty} \tag{1}$$

and its k-effective rank by

$$r_k(A) := \frac{\sum_{i=k+1}^p \lambda_i}{\lambda_{k+1}}. \tag{2}$$

Moreover, we define the orthogonal projectors $\pi_k(A) := \sum_{i=1}^k v_i v_i^\top$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ is the non-increasing sequence of eigenvalues of A. For given sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we say that $a_n = \Omega(b_n)$ (resp. $a_n = \mathscr{O}(b_n)$) if for some $c > 0$, $a_n \geq cb_n$ (resp. $a_n \leq cb_n$) for all integers n large enough. We will also write $a_n = o(b_n)$ or $a_n \ll b_n$ if $a_n/b_n \to 0$ as n goes to ∞.

1.2 Statement of the problem

Consider the two-component sub-Gaussian mixture model, where for $i = 1, \ldots, n$ we observe pairs (Y_i, η_i) such that

$$Y_i = \theta\eta_i + W_i. \tag{3}$$

Here, $\theta \in \mathbb{R}^p$ is a fixed but unknown vector, $\eta_i \in \{-1, 1\}$, $i = 1, \ldots, n$ are independent labels with uniform marginal distribution, and $W_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ are i.i.d sub-Gaussian vectors with covariance matrix Σ of full rank p. Moreover W_i and η_i are independent. More precisely, given the spectral decomposition $\Sigma = V\Lambda V^\top$, we assume that

$$W_i = V\Lambda^{1/2} w_i,$$

where w_i has components that are independent and 1-sub-Gaussian, meaning that for all $\lambda \in \mathbb{R}^p$,

$$\mathbf{E}(\exp(\lambda^\top w_1)) \leq \exp(\|\lambda\|^2/2).$$

Here, $\|\cdot\|$ denotes the Euclidean norm. We mostly focus on the supervised setting, where we are interested in constructing a classifier $\hat{\eta}$ from the training data

$$(\mathbf{Y}, \eta) = (Y_1, \eta_1), \ldots, (Y_n, \eta_n).$$

Let (Y_{n+1}, η_{n+1}) be an independent copy of (Y_1, η_1). The generalization error is defined via

$$\mathscr{R}_\Sigma(\hat{\eta}) := \mathbb{P}\left(\hat{\eta}((\mathbf{Y}, \eta); Y_{n+1}) \neq \eta_{n+1} | (\mathbf{Y}, \eta)\right),$$

both in probability and in expectation, where $\mathbb{P}$ is the probability under the model described above. Observe that

$$\mathbf{E}(\mathscr{R}_\Sigma(\hat{\eta})) = \mathbb{P}\left(\hat{\eta}((\mathbf{Y}, \eta); Y_{n+1}) \neq \eta_{n+1}\right).$$

When there is no ambiguity, we will omit the subscript Σ from $\mathscr{R}_\Sigma(\hat{\eta})$. We will primarily focus on the minimax risk

$$\inf_{\hat{\eta}} \sup_{\|\theta\|^2 \geq \Delta^2 \|\Sigma\|_\infty} \mathbf{E}(\mathscr{R}_\Sigma(\hat{\eta})),$$

where $\Delta > 0$ controls the degree of "separation" between the components of the mixture. Our main goal is to understand the precise dependence of the minimax risk on the key parameters of the problem, namely the dimension p, the sample size n, the effective rank of Σ, formally defined in display(1) below. This will allow us to deduce the necessary and sufficient conditions for consistent classification expressed in terms of parameter Δ. Here, consistency refers to vanishing minimax risk: we are interested in the necessary and sufficient conditions on the sequence $(\Delta_n)_{n\geq 1}$ such that

$$\inf_{\hat{\eta}} \sup_{\|\theta\|^2 \geq \Delta_n^2 \|\Sigma_n\|_\infty} \mathbf{E}(\mathscr{R}_\Sigma(\hat{\eta})) \underset{n\to\infty}{\to} 0,$$

where Δ_n, p_n and $\|\Sigma_n\|$ can change with n. Note that we are interested in the separation with respect to the Euclidean norm rather than the "Mahalanobis norm" given by

$$\|\theta\|_\Sigma^2 := \theta^\top \Sigma^{-1} \theta.$$

The Euclidean distance is often the preferred choice in estimation problems as its definition does not depend on the unknown covariance Σ but at the same time the error bounds can be adaptive to Σ, as will be the case in our framework.

The isotropic case $\Sigma = \mathbf{I}_p$, where the Euclidean and the Mahalanobis distances coincide, was previously investigated in Ndaoud (2022). In particular, it was shown that

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

1. We write $a_n \asymp b_n$ for two nonnegative sequences $\{a_n\}_{n\geq 1}$, $\{b_n\}_{n\geq 1}$ if $\frac{b_n}{c} \leq a_n \leq c b_n$ for some $c > 0$ and all $n \geq 1$.

When Σ is known and the noise W is normally distributed, it is easy to see that $\Sigma^{-1/2}Y$ follows the isotropic Gaussian mixture model where the signal vector is given by $\Sigma^{-1/2}\theta$. Following the reasoning similar to Ndaoud (2022), we can show in the general case that

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

In particular, the condition $\|\theta\|_\Sigma^2 \gg \sqrt{p/n}+1$ is necessary and sufficient for consistency under the norm $\|\cdot\|_\Sigma$. Moreover, the minimax optimal classifier is produced by the Linear Discriminant Analysis (LDA) and is given by

$$\hat{\eta}_{\mathrm{LDA}}(y) := \mathrm{sign}\left(\left\langle \Sigma^{-1}\left(\sum_{i=1}^n Y_i\eta_i\right), y\right\rangle\right). \tag{4}$$

The LDA estimator and its adaptive variants have been recently studied by several authors. Representative papers include Bing and Wegkamp (2022); Wang et al. (2020); Davis et al. (2021); Chen and Zhang (2021) and Cai and Zhang (2018). Most of these works consider anisotropic mixtures in the low-dimensional case. In this paper we are interested in the high-dimensional case ($p \gg n$), moreover, we assume that Σ is unknown. In this situation, the problem of estimating θ and Σ presents bigger challenges.

Another problem we are interested in is related to the risk $\mathscr{R}(\hat{\eta})$ of certain classifiers $\hat{\eta}$ of the form

$$\hat{\eta}(y) := \mathrm{sign}\left(\hat{\theta}^\top y\right),$$

where $\hat{\theta}$ is associated with a separating hyperplane. More precisely, we will focus on the minimum ℓ_2-norm solutions of the optimization problems corresponding either to the Ordinary Least Squares (OLS) or the Support Vector Machines (SVM). The hard margin SVM classifier $\hat{\theta}_{\mathrm{SVM}}$ is the solution to the problem

$$\hat{\theta}_{\mathrm{SVM}} = \arg\min_{\theta\in\mathbb{R}^p} \|\theta\|^2 \text{ subject to } \eta_i\theta^\top Y_i \geq 1, \quad \forall i = 1,\dots,n. \tag{5}$$

Similarly, $\hat{\theta}_{\mathrm{OLS}}$ is defined as the solution to the problem

$$\hat{\theta}_{\mathrm{OLS}} = \arg\min_{\theta\in\mathbb{R}^p} \|\theta\|^2 \text{ subject to } \eta_i\theta^\top Y_i = 1, \quad \forall i = 1,\dots,n. \tag{6}$$

While SVM is more commonly used for classification, both approaches have attracted significant attention, in part due to the discovery Hsu et al. (2020) of the phenomenon known as "Proliferation of Support Vectors" (SVP). Specifically, SVP corresponds to the situation when $\hat{\theta}_{\mathrm{SVM}} = \hat{\theta}_{\mathrm{OLS}}$, and typically occurs in the high-dimensional setting.

2. Related work and contribution

This section reviews key contributions to benign overfitting and its implications for high-dimensional classification while also outlining our own contributions.

2.1 Related work

Recent papers such as Belkin et al. (2018) and Belkin et al. (2019b) suggest that interpolating solutions (specifically, solutions that fit the training data perfectly) can achieve optimal rates for the problems of nonparametric regression and k-nearest neighbour clustering respectively. Termed "benign overfitting" by Bartlett et al. (2020), this phenomenon has been studied analytically in the framework of linear regression with isotropic noise. It was extended to the anisotropic case later in Wu and Xu (2020), where the authors proposed a fruitful idea of "alignment" and "misalignment" between the signal and the covariance matrix of the noise.

There is a number of recent works exploring the subject of overfitting in regression, while our focus is on the derivation of tight bounds for the misclassification rate in the framework of anisotropic sub-Gaussian mixture models. In moderate dimensions ($p \leq n$), this problem has been studied extensively. The recent works on the topic include Wang et al. (2020); Davis et al. (2021) and Chen and Zhang (2021), and efficient approaches such as perturbed gradient descent and modifications of Lloyd's algorithm have been proposed. When overparametrization (corresponding to the case $p \geq n$) is possible, support vector proliferation Muthukumar et al. (2020) has emerged as an important aspect of the analysis of interpolating SVM classifiers. In particular, papers including Ardeshir et al. (2021); Wang and Thrampoulidis (2020) establish sufficient conditions for the consistency of SVM interpolation. On the other hand, the work Liang and Recht (2021) approaches the topic in a more general setting via analyzing properties of Reproducing Kernel Hilbert Spaces (RKHS). Recent papers Cao et al. (2021) and Wang and Thrampoulidis (2020) are closely aligned with our work, but only consider the case when $r(\Sigma) = \Omega(n)$ and no regularization is present. We summarize their key findings in Table 1.

	Wang and Thrampoulidis (2020)	Cao et al. (2021)
SVP conditions	$\mathrm{Tr}(\Sigma) > C\left(\|\mathbf{\Sigma}\|_F \cdot n\sqrt{\log(n)} + \|\Sigma\|_\infty \cdot n^{3/2}\log(n)\right)$ and $\mathrm{Tr}(\Sigma) > C_1 n\sqrt{\log(n)\theta^\top\Sigma\theta}$	$\mathrm{Tr}(\Sigma) \geq C\max\left\{n^{3/2}\|\Sigma\|_\infty, n\|\Sigma\|_F\right\}$ and $\mathrm{Tr}(\Sigma) > C_1 n\sqrt{\log(n)\theta^\top\Sigma\theta}$
Error bounds	$\exp\left(\frac{-\left(\|\theta\|^2 - \frac{C_1 n\theta^\top\Sigma\theta}{\mathrm{Tr}(\Sigma)} - C_2\sqrt{\theta^\top\Sigma\theta}\right)^2}{C_3\max\left\{1, \frac{n^2\theta^\top\Sigma\theta}{\mathrm{Tr}(\Sigma)^2}\right\}\|\Sigma\|_F^2 + C_4\theta^\top\Sigma\theta}\right)$	$\exp\left(\frac{-C'\|\theta\|^4}{\theta^\top\Sigma\theta + \|\Sigma\|_F^2/n + \|\Sigma\|_\infty^2}\right)$

Table 1: Existing error bounds for the OLS interpolating classifier and the corresponding SVP conditions.

For the ease of interpreting the results given in table 1, let us consider the case $\Sigma = \mathbf{I}_p$. In this case, results of both works imply that the risk of the interpolating classifier is upper bounded by

$$\exp\left(-\frac{C'\|\theta\|^4}{p/n + \|\theta\|^2}\right),$$

where C' is a numerical constant. As we will demonstrate later, this result implies that the OLS interpolating solution $\hat{\eta}_{\mathrm{OLS}}$ is optimal in the case under consideration. In order

to extend their result to the SVM interpolating solution, both papers impose the condition $p = \Omega(n^{3/2})$, up to the logarithmic factors. This requirement is rather restrictive, as SVP phenomenon happens under the milder condition $p = \Omega(n \log(n))$ in other models. Authors of Wang and Thrampoulidis (2020) suggested that sharpening their requirement for SVP can be an interesting research question.

2.2 Contributions

Our main results are summarized below and compared with the previous state of the art.

- In section 3, we derive minimax bounds for the classification risk in the anisotropic sub-Gaussian mixture model, and show that the "averaging classifier", formally defined below, is adaptive and minimax optimal.

- In section 4, we study classification risk of a family of classifiers associated with solutions of regularized least squares problems. We deduce that, under mild assumptions, the interpolating solution $\hat{\eta}_{\mathrm{OLS}}$ is minimax optimal in the high-dimensional regime. Our results in this direction constitute an improvement over the existing bounds. We also expose interesting cases where the interpolating solution strictly outperforms the averaging classifier, which itself is closely related to the regularized classifiers, under mild assumptions on the covariance structure of the noise.

- In section 5, we show that the support vector proliferation phenomenon occurs under the mild conditions, namely, $r(\Sigma) = \Omega(n \log(n))$ and $\mathrm{Tr}(\Sigma) = \Omega(n\sqrt{\log(n)\theta^\top \Sigma \theta})$. As a consequence, we deduce the risk bounds for the hard-margin SVM classifier. The aforementioned conditions are strictly better than the previously known ones, as the latter require that $r(\Sigma) = \Omega(n^{3/2} \log(n))$.

- Finally, in section 6, we propose the outlier contamination framework where interpolation leads to robust minimax optimal classifiers over large classes of signal vectors, while both the averaging and the LDA estimators fail. We conclude that not only interpolation is benign, but it can also lead to optimality and robustness.

3. Minimax rates for classification

In this section we consider the supervised classification problem with sub-Gaussian noise with unknown covariance Σ. In order to do so, let us define $\mathscr{P}_{\mathrm{G}}(\Delta, r)$ to be the class of distributions of the pair (Y, η) where $Y = \eta\theta + \Sigma^{1/2} w$, η is a random sign and $w \sim \mathcal{N}(0, \mathbf{I}_p)$. Moreover, the mean and covariance of the marginal distribution of Y satisfy the inequalities

$$r(\Sigma^2) = r, \ \|\Sigma\|_\infty > 0, \ \|\theta\|^2 \geq \Delta^2 \|\Sigma\|_\infty.$$

The class $\mathscr{P}_{\mathrm{SG}}(\Delta, r)$ is defined similarly, with the only difference being that w can have a sub-Gaussian distribution that we defined in section 1.2, instead of only Gaussian. The following theorem presents the minimax lower bound over the class $\mathscr{P}_{\mathrm{G}}(\Delta, r)$ (hence, it automatically holds over the larger class $\mathscr{P}_{\mathrm{SG}}(\Delta, r)$).

Theorem 1 *Let $\Delta, r > 0$. Then*

$$\inf_{\hat{\eta}} \sup_{\mathscr{P}_{\mathrm{G}}(\Delta,r)} \mathbf{E}(\mathscr{R}_{\Sigma}(\hat{\eta})) \geq C \exp\left(-c \frac{\Delta^4}{\Delta^2 + \frac{r}{n}}\right),$$

for some absolute constants $c, C > 0$ and where the infimum is taken over all measurable classifiers based on $\{Y_i, \eta_i\}_{i=1}^n$.

Notice that the inequality is stated in terms of the Euclidean norm of θ, and not the Mahalanobis norm that would lead to a different lower bound. This particular choice is motivated by the fact that the LDA classifier, that is minimax optimal under the Mahalanobis norm separation, is not adaptive as it requires prior knowledge of Σ. Here, we are seeking adaptive classifiers that are optimal over large classes of Σ which is typically unknown in applications. The proof of the lower bound is inspired by the argument in Ndaoud (2022) that only holds for the isotropic noise. As for the upper bound, we show that the "averaging" linear classifier defined as

$$\hat{\eta}_{\mathrm{ave}}(y) := \mathrm{sign}\left(\left\langle \hat{\theta}_{ave}, y \right\rangle\right),$$

where $\hat{\theta}_{ave} = \frac{1}{n}\sum_{i=1}^n Y_i \eta_i$, is minimax optimal. More specifically, we will prove the following inequality.

Theorem 2 *Assume that (Y_i, η_i), $i = 1, \ldots, n$ are distributed according to model (3). Then, with probability at least $1 - \delta - e^{-c_1 n}$,*

$$\mathscr{R}(\hat{\eta}_{ave}) \leq C \exp\left(-c_2 \frac{\|\theta\|^4}{\theta^\top \Sigma \theta + \frac{\mathrm{Tr}(\Sigma^2) + \|\Sigma\|_\infty^2 \log(1/\delta)}{n}}\right).$$

Moreover, for any $\Delta, r > 0$,

$$\sup_{\mathscr{P}_{\mathrm{SG}}(\Delta,r)} \mathbf{E}(\mathscr{R}(\hat{\eta}_{ave})) \leq C \exp\left(-c_3 \frac{\Delta^4}{\Delta^2 + \frac{r}{n}}\right).$$

Here, $c_\cdot, c_2, c_3$ and C are positive absolute constants.

Theorem 1 and Theorem 2 together imply that a necessary and sufficient condition, in the minimax sense, for consistency under the Euclidean norm is given by

$$\|\theta\|^2 \gg \|\Sigma\|_\infty (\sqrt{r(\Sigma^2)/n} + 1).$$

In order to complete our result, we state here a corresponding lower bound for the averaging classifier that shows that our pointwise upper bound can not be improved in general.

Proposition 3.1 *Let $n, r(\Sigma^2)$ be large enough. Then for any $\theta \in \mathbb{R}^p$*

$$\mathbf{E}(\mathscr{R}(\hat{\eta}_{ave})) \geq C' \exp\left(-c' \frac{\|\theta\|^4}{\theta^\top \Sigma \theta + \mathrm{Tr}(\Sigma^2)/n}\right),$$

for some absolute constants $c', C' > 0$.

One way to understand superiority of the averaging classifier in this case is to observe that the price of estimating θ under the Euclidean norm, depends on $r(\Sigma)$ while it depends on p under the Mahalanobis norm. This fact is stated explicitly next.

Proposition 3.2 *Let $p \geq n$. Then for any $\theta \in \mathbb{R}^p$*

$$\mathbf{E}(\mathscr{R}(\hat{\eta}_{LDA})) \geq C \exp\left(-c \frac{\|\theta\|_\Sigma^4}{\|\theta\|_\Sigma^2 + \frac{p}{n}}\right),$$

for some $c, C > 0$.

Among other conclusions, this statement implies that from a minimax perspective, the averaging classifier outperforms the LDA . This phenomenon is surprising as the LDA classifier is nothing but the Bayes classifier with θ replaced by its unbiased estimator. We remark that the following upper bound for the risk $\mathscr{R}(\hat{\eta}_{\mathrm{LDA}})$ can be deduced immediately from Theorem 2 applied to the "whitened" data $\Sigma^{-1/2}Y$: in this case, we simply need to replace θ by $\Sigma^{-1/2}\theta$ and Σ by I_p, leading to the inequality

$$\mathscr{R}(\hat{\eta}_{\mathrm{LDA}}) \leq C \exp\left(-c_2 \frac{\|\theta\|_\Sigma^4}{\|\theta\|_\Sigma^2 + \frac{p+\log(1/\delta)}{n}}\right)$$

that holds with probability at least $1 - \delta - e^{-c_1 n}$. Observe that in the worst case scenario, the vector θ is chosen so that $\Sigma\theta = \|\Sigma\|_\infty \theta$, whence the upper bound for the risk of $\hat{\eta}_{\mathrm{ave}}$ is smaller than the lower bound for $\mathscr{R}(\hat{\eta}_{\mathrm{LDA}})$, implying its sub-optimality. This phenomenon is only possible in high dimensions. In this worst case scenario, consistency of the averaging classifier requires that $\|\theta\|^2 = \Omega(\|\Sigma\|_\infty(1 + \sqrt{r(\Sigma^2)/n}))$ while for the LDA classifier, the requirement is $\|\theta\|^2 = \Omega(\|\Sigma\|_\infty(1 + \sqrt{p/n}))$. Let us emphasize here that our choice of separation under the Euclidean norm leads to conditions that depend on the "intrinsic" dimension of the problem expressed in terms of $r(\Sigma)$, while under the Mahalanobis norm the optimal separation threshold depends on dimension p.

The picture is rather different in low dimensions. It is in fact easy to see that when $p \ll n$, LDA outperforms the averaging classifier for any given vector θ whenever consistent classification is possible (i.e. $\|\theta\|^2/\|\Sigma\|_\infty \to \infty$). Indeed, in this case

$$\frac{\|\theta\|_\Sigma^4}{\|\theta\|_\Sigma^2 + \frac{p}{n}} = \Omega\left(\|\theta\|_\Sigma^2\right).$$

Moreover, it is always true that

$$\|\theta\|_\Sigma^2 \geq \frac{\|\theta\|^4}{\theta^\top \Sigma \theta}.$$

Hence, whenever consistency is achievable and whenever $p \ll n$, we have that

$$\frac{\|\theta\|_\Sigma^4}{\|\theta\|_\Sigma^2 + \frac{p}{n}} = \Omega\left(\frac{\|\theta\|^4}{\theta^\top \Sigma \theta + \frac{\mathrm{Tr}\,(\Sigma^2)}{n}}\right).$$

Notice that the last statement is stronger than minimax comparison. We conclude that the LDA classifier always outperforms the averaging classifier in low dimensions.

4. Interpolation versus regularization in the Gaussian mixture framework

In this section, we study the risk of the regularized OLS estimators. While it is more common to study SVM for classification, recent works Ardeshir et al. (2021); Hsu et al. (2020) have shown that in high dimensions (specifically, $p = \Omega(n \log n)$), SVM and OLS solutions coincide under mild conditions. This phenomenon is known in the literature as the "proliferation of support vectors". One of its implications is that in high dimensions, it is sufficient to study the properties of the least squares estimator and then demonstrate that it coincides with the hard-margin SVM. For the rest of this section, our goal is to study the risk of the family of supervised estimators defined as solutions to the problem

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

Observe that the case $\lambda \to 0$ and $p \geq n$ leads to interpolation, specifically to the minimum ℓ_2-norm interpolating solution, or equivalently $\hat{\theta}_0 = \hat{\theta}_{\mathrm{OLS}}$. For each $\lambda > 0$, the corresponding estimator $\hat{\theta}_\lambda$ is proportional to

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

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

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

In what follows, we will derive upper bounds for the risk $\mathscr{R}(\hat{\eta}_\lambda)$. We will also provide sufficient conditions for the matrix Σ ensuring that the interpolating classifier (corresponding to $\lambda = 0$) achieves at least the same performance as the averaging classifier that is minimax optimal. Notice that $\lambda \hat{\theta}_\lambda \underset{\lambda \to \infty}{\to} \hat{\theta}_{ave}$ (where $\hat{\theta}_{ave} = \frac{1}{n} \sum_{i=1}^{n} Y_i \eta_i$) and hence we recover the averaging classifier as $\lambda \to \infty$. Indeed, it is true that

$$\frac{\hat{\theta}_\lambda}{\|\hat{\theta}_\lambda\|} = \frac{\lambda \hat{\theta}_\lambda}{\|\lambda \hat{\theta}_\lambda\|} \to \frac{\hat{\theta}_{ave}}{\|\hat{\theta}_{ave}\|} \text{ as } \lambda \to \infty.$$

Since the classifier $\hat{\eta}_\lambda$ only depends on the direction $\frac{\hat{\theta}_\lambda}{\|\hat{\theta}_\lambda\|}$ of $\hat{\theta}_\lambda$, it follows that $\hat{\eta}_\lambda$ converges to $\hat{\eta}_{ave}$ as $\lambda \to \infty$. We emphasize here that if we replace the regularization term $\|\theta\|^2$ by $\|\theta\|_\Sigma^2$, then our claims may no longer be true.

As a side note, we conjecture that the excess risk of estimating θ itself gets smaller for large values of λ although this is not necessarily the case for the classification risk. It is therefore reasonable to suggest that the estimation risk associated to the parameter θ may not be the right metric to evaluate classification performance.

Recall that $r_k(A) = \frac{\sum_{i=k+1}^{p} \lambda_i}{\lambda_{k+1}}$. In order to state the main result of this section, we will need to introduce the following quantity. For a given covariance matrix Σ, we define the function k_Σ^* via

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

Here, $C_1 > 1$ is a sufficiently large absolute constant and $k^*_\Sigma(\lambda) := p+1$ if the above set is empty. The reader may observe that $k^*_\Sigma(\cdot)$ is decreasing with λ and that $k^*_\Sigma(\lambda) = 0$ if $r(\Sigma) \geq C_1 n$. In what follows, we will require $k^*_\Sigma(\lambda)$ to be smaller than $n/2$. For $\lambda = 0$, this means that we are not allowing more than a fraction of all eigenvalues to be much larger than the remainder of the spectrum. This condition encompasses many covariance structures of interest, and in particular covers the case when $\Sigma = \mathbf{I}_p + R$ where R can be viewed as a low rank perturbation/corruption component. In what follows, π_{k^*} will stand for the orthogonal projection $\pi_{k^*}(\Sigma)$.

Theorem 3 *Let $\Delta > 0, \lambda \geq 0$. Assume that $k^*_\Sigma(\lambda) \leq n/2$ and that $\|\pi_{k^*}\theta\|^2 \leq \|\theta\|^2/5$. Then for some absolute constants $c, C > 0$,*

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

with probability at least $1 - \delta - e^{-cn}$, where $k^* = k^*_\Sigma(\lambda)$.

Remark 4

1. *The condition $\|\pi_{k^*}\theta\|^2 \leq \|\theta\|^2/5$ connecting θ with Σ can be understood as follows. One may think of the k^* eigenvectors corresponding to largest eigenvalues of Σ as "outliers," or directions affecting dramatically the rest of the spectrum. Our condition prohibits θ from concentrating too much of its mass in the subspace spanned by the latter eigenvectors. For instance, if $\theta/\|\theta\|$ is a random vector with uniform distribution on the unit sphere then $\|\pi_{k^*}\theta\| \approx \frac{k^*}{p}\|\theta\| \ll \|\theta\|$. The same remark also holds if we fix θ and think of the range of π_{k_*} as being a random subspace uniformly distributed over the Grassmannian, thus making $\pi_{k_*}\theta$ a random vector. In other words, the condition $\|\pi_{k^*}\theta\|^2 \leq \|\theta\|^2/5$ simply means that the vector θ is only allowed to be aligned with the "clean" part of the spectrum of the covariance Σ.*

2. *When $k^*_\Sigma(\lambda) = 0$, or equivalently $r(\Sigma) \geq C_1 n$, we recover the bound obtained in Cao et al. (2021) by taking $\delta = e^{-cn}$. Notice that in this case the "alignment condition" $\|\pi_{k^*}\theta\|^2 \leq \|\theta\|^2/5$ is always satisfied, as $\pi_{k^*} = 0_p$. Our result is stronger since we show that the bound holds with probability $1 - e^{-cn}$ while in Cao et al. (2021) authors only prove that the same bound holds with probability $1 - 1/n$.*

3. *Whenever $r(\Sigma) \geq C_1 n$ and under the additional mild assumption $r(\Sigma^2) \geq \log(n)$, Theorem 4.1 with the choice $\delta = 1/n$ implies that*

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

with probability at least $1 - 1/n$. Therefore, taking the limit as $\lambda \to 0$ (interpolation) leads to the same bound as the averaging classifier ($\lambda \to \infty$) in this case. The risk bound for the latter is given in Theorem 2). In other words, all regularized classifiers have similar performance whenever $r(\Sigma) \geq c_1 n$.

In the case of moderate values of k^*, recalling that $k^*(\cdot)$ is a decreasing function of λ, we see that $k^*\lambda_{k*}^2 + \sum_{i>k^*} \lambda_i^2$ is an increasing function of λ such that

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

The quantity $k^*\lambda_{k*}^2 + \sum_{i>k^*} \lambda_i^2$ could be viewed as a truncated trace of Σ^2, where truncation is applied to the large eigenvalues of Σ. Unlike the previous works Cao et al. (2021); Wang and Thrampoulidis (2020), our result is more general since we allow k^* to be non-zero. The upper bound of Theorem 3 in particular gets smaller as λ goes to 0, which indicates that interpolation (corresponding to $\lambda = 0$) may outperform regularization for large values of λ (corresponding to the averaging classifier in the limit $\lambda \to \infty$), in particular in the case where a small number of eigenvalues of Σ are much larger than the rest of the spectrum. To illustrate this fact, let us present a concrete example.

Example 1 *Assume that $p \geq 2C_1 n$ and let Σ be a covariance matrix with $\lambda_i(\Sigma) = 1 + 2^{r-i}\mu$ for $i = 1, \ldots, r$ and $\lambda_i(\Sigma) = 1$ for $i = r, \ldots, p$. We choose the parameters r and μ such that $r \leq C_1 n/4$ and $\mu \geq 4p/(C_1 n)$. In this case,*

$$\mathrm{Tr}(\Sigma) = p + \mu(2^r - 1), \text{ and}$$
$$r_k(\Sigma) = \frac{p - k + \mu(2^{r-k} - 1)}{1 + 2^{r-k-1}\mu}$$

for any $k < r$, and $r_k(\Sigma) = p - k$ for any $k \geq r$. On the one hand, we see that $k^* := k_\Sigma^*(0) = r$ and that

$$k^*\lambda_{k^*}^2 + \sum_{i>k^*} \lambda_i^2 = r(1+\mu)^2 + (p - r)$$

which is of order $r\mu^2 + p$. On the other hand, we see that

$$\mathrm{Tr}(\Sigma^2) = p + \mu^2 \frac{4^r - 1}{3} + 2\mu(2^r - 1),$$

which is of order $4^r\mu^2 + p$. Hence, as r grows, we conclude that $\mathrm{Tr}(\Sigma^2) \gg k^*\lambda_{k^*}^2 + \sum_{i>k^*} \lambda_i^2$. Observe that if θ is chosen so that $\Sigma\theta = \|\Sigma\|_\infty \theta$, Proposition 3.1 implies that the averaging classifier requires $\|\theta\|^2 = \Omega\left(2^r\mu(1 + \sqrt{\frac{4^r\mu^2+p}{n}})\right)$ to guarantee consistency. On the other hand, it suffices that

$$\|\theta\|^2 = \Omega\left(2^r\mu\left(r + \sqrt{\frac{r\mu^2 + p}{n}}\right)\right)$$

for the interpolating classifier, in view of Theorem 3. If r is chosen so that $2^r\mu \gg \max(r\sqrt{n}, \sqrt{p})$, we deduce that the interpolating classifier is strictly better.

In the case when k^* is large, we get the following result without any further assumptions on Σ. Let us define

$$\mathscr{C}_{k^*}(\Sigma) := \{\theta \in \mathbb{R}^p, \quad \|\pi_{k^*}\theta\| \leq \|\theta\|/\sqrt{5}\}. \tag{7}$$

Proposition 4.1 *Let $\Delta > 0, \lambda \geq 0$. Assume that $k^*_\Sigma(\lambda) \leq n/2$. Then there exist absolute constants $c, C > 0$ such that*

$$\sup_{\substack{\|\theta\|^2 \geq \Delta^2 \|\Sigma\|_\infty \\ \theta \in \mathscr{C}_{k^*}(\Sigma)}} \mathbf{E}(\mathscr{R}(\hat{\eta}_\lambda)) \leq C \exp\left(-c \frac{\Delta^4}{\Delta^2 + \frac{r(\Sigma^2)}{n}}\right) + e^{-cn}.$$

This result suggests that under a mild condition on the covariance of the noise, not only interpolation is benign but it is also minimax optimal on the set $\mathscr{C}_{k^*}(\Sigma)$. This also means that interpolation is better for classification than for regression since it does not suffer from a bias term which often leads to bad worst-case performance Bartlett et al. (2020).

5. Proliferation of support vectors in high dimensions under the sub-Gaussian mixture model

In this section, we provide sufficient conditions for proliferation of support vectors. Based on the results in Hsu et al. (2020), $\hat{\theta}_{\mathrm{SVM}}$ and $\hat{\theta}_{\mathrm{OLS}}$, as defined in display (5)-(6), 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. In the remainder of the section we denote $k^* := k^*_\Sigma(0)$. The main result is stated next.

Theorem 5 *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 $\sqrt{\theta^\top \Sigma \theta (1+k^*)\log(n)} \leq C \sum_{i>k^*} \lambda_i / n$ for some absolute constant $C > 0$. Then with probability at least $1 - 1/n$,*

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

As a consequence, $\hat{\eta}_{\mathrm{SVM}}$ attains the same performance as $\hat{\eta}_0$ under the conditions of Theorem 5. When $k^* = 0$ (i.e. $r(\Sigma) = \Omega(n)$), sufficient conditions read as

- $\sqrt{\theta^\top \Sigma \theta \log(n)} \leq C \operatorname{Tr}(\Sigma)/n$;
- $\operatorname{Tr}(\Sigma^2) n \log(n) \leq C(\operatorname{Tr}(\Sigma))^2$.

The first condition (that is signal-dependent) is also required in prior works Cao et al. (2021); Wang and Thrampoulidis (2020). As for the "dimension-dependent" second condition, it is much milder than the one proposed in the previous papers on the topic. To compare these results, consider the case $\Sigma = \mathbf{I}_p$, where our condition requires that $p = \Omega(n \log(n))$ while the earlier known versions need $p = \Omega(n^{3/2} \log(n))$. Our result also supports the claim that $r(\Sigma) = \Omega(n \log(n))$ suffices for proliferation under the sub-Gaussian mixture model, which is the general conjecture stated in Hsu et al. (2020).

6. Application to robust supervised classification

In this section, we present the framework for robust classification in the Gaussian mixture model. In the rest of this section we consider the case of Gaussian noise with identity covariance $\Sigma = \mathbf{I}_p$. We will assume that the training set has a different covariance than the

covariance of the new observation to be classified, due to the action of a malicious adversary. More precisely, given the vector of observations Y, the adversary can corrupt the sample as follows: she chooses up to $r \leq n/4$ eigenvalues of the covariance matrix Σ and positive scalars $O_1, \dots, O_r$, and then adds i.i.d. random noise to Y such that the new observations become

$$\tilde{Y}_i = Y_i + O^{1/2}\varepsilon_i = \eta_i\theta + W_i + O^{1/2}\varepsilon_i, \quad \forall i = 1, \dots, n,$$

where ε_i are i.i.d. standard normal vectors that are also independent from Y, and $O = \sum_{i\in R} O_i e_i e_i^\top$ where $\{e_1, \dots, e_p\}$ is the canonical Euclidean basis of $\mathbb{R}^p$ and R is the set of indices corresponding to the corrupted eigenvalues. Observe that the covariance of the noise is now given by $\mathbf{I}_p + O$. In what follows, π_r denotes the projection $\pi_r(O) := \sum_{i\in R} e_i e_i^\top$. Our goal is to show that under minimal assumptions, the interpolating classifier $\hat{\eta}_0$ is still minimax optimal while both the averaging and the LDA classifiers fail to perform well. Recall the definition (7) of the set $\mathscr{C}_r(\mathbf{I}_p + O)$.

Theorem 6 *Assume that $r \leq n/4$ and that $\Delta^2 \geq p/n$. Then*

$$\sup_{\substack{\|\theta\|\geq\Delta \\ \theta\in\mathscr{C}_r(\mathbf{I}_p+O)}} \mathbf{E}(\mathscr{R}(\hat{\eta}_0)) \leq C\exp\left(-c\frac{\Delta^4}{\Delta^2 + \frac{p}{n}}\right) + e^{-cn}.$$

One implication of this theorem is the fact that $\hat{\eta}_0$ is minimax optimal on the set $\mathscr{C}_r(\mathbf{I}_p + O)$ and robust with respect to corruption, for the moderate values of p. When $r \leq n/4$ and the direction of θ is not too closely aligned with the eigenvectors corresponding to the corrupted part of the spectrum, then $\hat{\eta}_0$ mimicks the performance of the averaging estimator in the absence of outliers. In order to compare the bound stated above to the bounds available for other estimators, we rely on the next proposition.

Proposition 6.1 *Assume that the noise is Gaussian and that O satisfies $O = n\sum_{i=1}^r e_i e_i^\top$ where $r = n/4$. Then for any θ such that $\|\theta\|^2 \leq \sqrt{n}$,*

$$\mathbf{E}(\mathscr{R}_\Sigma(\hat{\eta}_{LDA})) \wedge \mathbf{E}(\mathscr{R}_\Sigma(\hat{\eta}_{ave})) \geq C$$

for some absolute constant $C > 0$.

In summary, when $r \leq n/4$, there exists a regime (namely, when $p/n \leq \|\theta\|^2 \leq \sqrt{n}$) such that interpolation $\hat{\eta}_0$ is minimax optimal over a large class of vectors θ, under the contamination model we presented, while both $\hat{\eta}_{ave}$ and $\hat{\eta}_{\mathrm{LDA}}$ can fail with constant non-zero probability.

7. Numerical experiments

The goal of this section is to compare the performance of several estimators $\hat{\eta}_\lambda$ for different values of λ. The case $\lambda = 0$ corresponds to interpolation, while $\lambda = \infty$ recovers the averaging classifier $\hat{\eta}_{\mathrm{ave}}$. In all our simulations we will only consider Gaussian noise and the situation when $\theta/\|\theta\|$ has uniform distribution over the unit sphere.

Our simulation setup is defined as follows. We choose $p = 500$ and $n = 30$ to allow for overparametrization. We compare 4 estimators: the interpolating classifier $\hat{\eta}_0$, classifiers

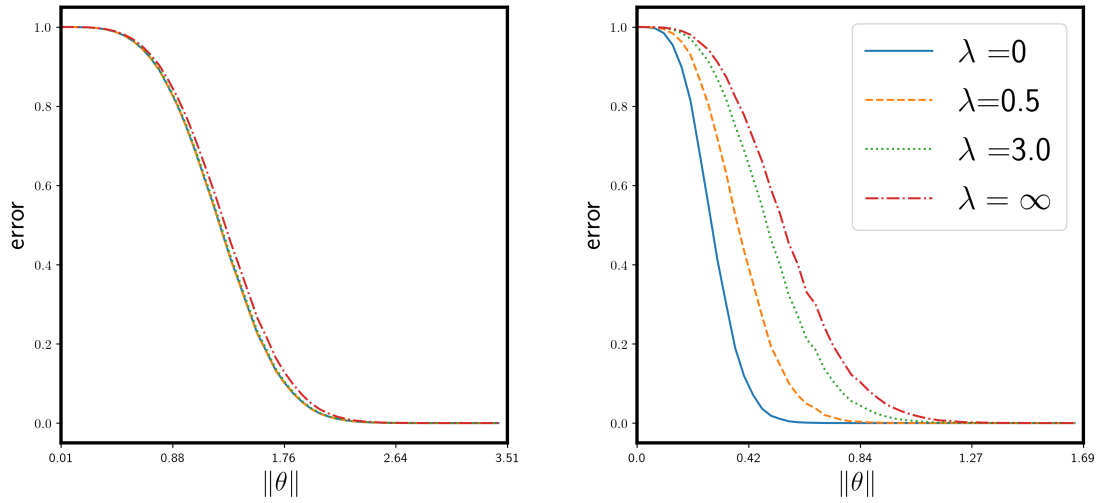

Figure 1: Comparison of the generalization error of four classifiers. The left plot corresponds to the case of large effective rank while the right one corresponds to medium effective rank.

$\hat{\eta}_\lambda$ for $\lambda = 0.5$ and $\lambda = 3$, as well as the averaging estimator $\hat{\eta}_{\text{ave}}$. For each value of $\|\theta\|$, simulation was repeated 1000 times; finally, we plot the empirical generalization error.

For our first experiment (Figure 1), we compare performance of the four classifiers in two cases:

- The case of large effective rank where we choose Σ to be a diagonal matrix with $\lambda_i = (p - i + 1)/p$ for all $i = 1, \dots, p$. This case corresponds to $k^*(0) = 0$;
- The case of medium effective rank where we choose Σ to be a diagonal matrix with $\lambda_1 = \lambda_2 = \lambda_3 = 1$ and $\lambda_4 = \cdots = \lambda_p = 0.01$. This case corresponds to $k^*(0) = 3$.

Consistent with the theoretical predictions, our simulations suggest that all classifiers have similar performance in the regime of large effective rank. Interestingly, interpolation seems to perform best when the effective rank is smaller than n. This confirms our observation that interpolation can be superior to regularization in some cases.

As for our second experiment (Figure 2), we choose $\Sigma = \mathbf{I}_p$ and we corrupt the training sample, as explained in Section 6, by setting randomly $n/2$ diagonal entries of the covariance to 1000. Remember that this modification only impacts the training sample while the test sample has isotropic noise.

In this case, the averaging classifier fails to predict new labels correctly while the interpolating classifier is still able to classify well despite the corruption. Observe also that regularized classifiers (corresponding to the small regularization parameter) perform similar to the interpolating classifier. This occurs simply due to the fact that $k^*_{\tilde{\Sigma}}(\lambda) = k^*_{\tilde{\Sigma}}(0)$ for all values of λ much smaller than the magnitude of the corruption, where $\tilde{\Sigma}$ is the corrupted covariance matrix. We conclude by reaffirming the claim that interpolation is not only

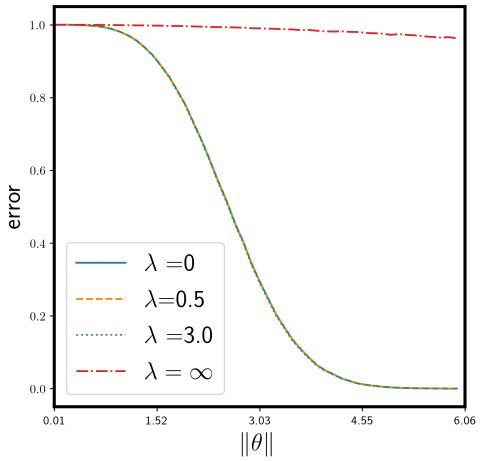

Figure 2: Comparison of the performance of four classifiers under corruption.

harmless in high dimensions, but can outperform regularization and in some cases exhibits an unexpected feature of robustness.

8. Future directions

In summary, our findings establish that in high-dimensional scenarios, the interpolating classifier attains minimax optimality and has the potential to surpass regularized classifiers. This holds true under mild assumptions on the covariance structure of the noise. Additionally, we presented a scenario wherein interpolation exhibits robustness to corruption while other classifiers do not.

Our analysis for the interpolating solution hinges on the SVP property. Unfortunately, this phenomenon only occurs under strong conditions on the separation, i.e. $\sqrt{\theta^\top \Sigma \theta} = \mathscr{O}(p/n)$, up to the logarithmic factors. While it is clear that SVP becomes harder to achieve as the separation gets larger, it is natural to expect that the SVM interpolating solution will still do well, especially under large separation. A particularly interesting direction to explore would be to prove similar results under large separation of the mixtures. In order to do so, we would need novel proof techniques beyond the ideas based on the SVP phenomenon. Another direction worth exploring is related to robustness. While our contamination setup is rather simple, it would be interesting to show that SVM is robust under more general corruption models.

Acknowledgments

S. Minsker acknowledges support by the National Science Foundation grants CIF-1908905 and DMS CAREER-2045068. M. Ndaoud acknowledges support by the National Science Foundation grant CIF-1908905.

Appendix A. Proofs of minimax bounds for classification

A.1 Proof of Theorem 1

Fix $\lambda, r > 0$ and let Σ be a diagonal positive semidefinite (PSD) matrix such that $r(\Sigma^2) = r$ and $\|\Sigma\|_\infty = \lambda$. Then

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

Therefore, we can focus on establishing the result for the diagonal matrix Σ. We will use the fact that

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

for any prior π on $(\theta, \eta, \eta_{n+1})$ such that $\|\theta\|^2 \geq \Delta^2\lambda$. The quantity $\mathbf{E}_{(\theta,\eta,\eta_{n+1})}(\cdot)$ stands for the expectation corresponding to $(Y; Y_{n+1})$ following model (3). The proof is decomposed into two steps.

- **A dimension-independent lower bound:** Let $\bar\theta$ be a fixed vector in $\mathbb{R}^p$ such that $\|\bar\theta\|^2 = \Delta^2\lambda$. We place independent Rademacher priors π_i on each η_i for $i = 1, \dots, n+1$ (that is, η_i's are random signs). It follows that

$$\inf_{\tilde\eta} \mathbb{E}_\pi \mathbf{E}_{(\bar\theta,\eta,\eta_{n+1})} |\tilde\eta((Y,\eta); Y_{n+1}) - \eta_{n+1}| \geq \inf_{\bar\eta} \mathbb{E}_{\pi_{n+1}} \mathbf{E}_{(\bar\theta,\eta_{n+1})} |\bar\eta(Y_{n+1}) - \eta_{n+1}|, \quad (8)$$

  where $\bar\eta(Y_{n+1}) = \mathbf{E}(\tilde\eta((Y,\eta); Y_{n+1})|Y_{n+1}) \in [-1, 1]$ as we average over (Y, η). The last inequality holds in view of Jensen's inequality and the independence between (Y, η) and η_{n+1}. We define, for $\epsilon \in \{-1, 1\}$, $\tilde f_\epsilon(\cdot)$ 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 optimal as it achieves the minimum of the RHS of (8).

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

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

  It follows that

$$\begin{aligned} \frac{\tilde f_1(Y_{n+1})}{\tilde f_{-1}(Y_{n+1})} &= \frac{(2\pi)^{-p/2}|\Sigma|^{-1/2} e^{-\frac{1}{2}(Y_{n+1}-\bar\theta)^\top\Sigma^{-1}(Y_{n+1}-\bar\theta)}}{(2\pi)^{-p/2}|\Sigma|^{-1/2} e^{-\frac{1}{2}(Y_{n+1}+\bar\theta)^\top\Sigma^{-1}(Y_{n+1}+\bar\theta)}} \\ &= e^{2\bar\theta^\top\Sigma^{-1}Y_{n+1}} \end{aligned}$$

  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 (8), we get further that

$$\inf_{\bar{\eta}} \mathbf{E}_\pi |\bar{\eta}(Y_{n+1}) - \eta_{n+1}| = 2\mathbf{E}\mathscr{R}_\Sigma(\eta^*).$$

It is straightforward to see that

$$\mathbf{E}(\mathscr{R}_\Sigma(\eta^*)) = \Phi^c\left(\sqrt{\bar{\theta}^\top \Sigma^{-1}\bar{\theta}}\right) \geq Ce^{-c\bar{\theta}^\top\Sigma^{-1}\bar{\theta}},$$

for some $c, C > 0$ where $\Phi(\cdot)$ is the standard normal cumulative function. The last inequality holds for any $\bar{\theta}$ 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

$$\inf_{\hat{\eta}} \sup_{\|\theta\|^2 \geq \Delta^2\lambda} \mathbf{E}(\mathscr{R}_\Sigma(\hat{\eta})) \geq Ce^{-c\Delta^2}.$$

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

$$\inf_{\hat{\eta}} \sup_{\|\theta\|^2 \geq \Delta^2\lambda} \mathbf{E}(\mathscr{R}_\Sigma(\hat{\eta})) \geq Ce^{-cn\Delta^4/r}.$$

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

- **A dimension-dependent lower bound:**

  Since Σ is diagonal, we write $\Sigma = \mathrm{diag}(d_1, \ldots, d_p)$ where $\lambda = d_1 \geq d_2 \geq \cdots \geq d_p > 0$. According to Theorem 1 in Ndaoud (2022) we have

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

  for any prior π on $(\theta, \eta, \eta_{n+1})$. The second term in the above lower bound accounts for the constraint on θ. In what follows, we fix η and choose π^D to be a product prior on (θ,η_{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 Rudelson and Vershynin (2013b), we deduce that

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

  for some $c, C > 0$. Since $\Delta^2 \leq 10r/n$, we only need to show that, for n large enough, we have

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

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

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

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

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

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

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

Hence

$$\begin{aligned}
\frac{\tilde{f}_1(Y)}{\tilde{f}_{-1}(Y)} &= \prod_{j=1}^{p} e^{-\frac{1}{2}L_j^\top((\Sigma_1^j)^{-1}-(\Sigma_{-1}^j)^{-1})L_j} \\
&= \prod_{j=1}^{p}\exp\left(\frac{2\Delta^2\lambda/\sum_i d_i^2}{1+(2n\Delta^2 d_j)/(\lambda r)}L_{j,n+1}\sum_{k=1}^{n}L_{jk}\eta_k\right) \\
&= \exp\left(\frac{2\Delta^2\lambda}{\sum_i d_i^2}\sum_{k=1}^{n}\eta_k\sum_{j=1}^{p}\frac{L_{jk}L_{j,n+1}}{1+(2n\Delta^2 d_j)/(\lambda r)}\right) \\
&= \exp\left(\frac{2\Delta^2\lambda}{\sum_i d_i^2}\langle Y_{n+1},\sum_{k=1}^{n}\eta_k\tilde{D}Y_k\rangle\right),
\end{aligned}$$

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

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

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

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

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 is clear that

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

for some $c, C > 0$. Therefore

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

Consider three events

$$\mathbb{A}_1 = \{\|\theta\|^2 \leq 2\Delta^2\lambda\},$$
$$\mathbb{A}_2 = \{\xi^\top\Sigma\xi \geq r\lambda^2/2n\},$$
$$\mathbb{A}_3 = \{\langle\theta, \tilde{D}\xi\rangle^2 \leq 2\Delta^4\lambda^2\}.$$

Since $\Delta^4 \geq \Delta^2$ by assumption, we get

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

Using Hanson-Wright inequality, we deduce that

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

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

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

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

A.2 Proof of Theorem 2

Let θ be a vector in $\mathbb{R}^p$. Without loss of generality we may assume that $\|\theta\|^4 \geq C_1\theta^\top\Sigma\theta$ for some constant $C_1 > 0$ large enough, otherwise the result is trivial as the upper bound becomes of constant order. For the rest of the proof, we use the notation $\hat{\eta} := \hat{\eta}_{\text{ave}}$. We start by observing that

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

Let us denote $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n Y_i\eta_i = \theta + \xi$ where $\xi = \frac{1}{n}\sum_{i=1}^n W_i\eta_i$. We get the following upper bound

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

where we have conditioned on $\hat{\theta}$ to get the last inequality and used the fact that $\eta_{n+1}w_{n+1}$ has also i.i.d 1-sub-Gaussian entries. 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

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

Let us define now the event

$$\mathscr{A} = \{\xi^\top \Sigma \xi \leq 3/2 \operatorname{Tr}(\Sigma^2)/n + \|\Sigma\|_\infty^2 \log(1/\delta)/n\} \cap \{4\langle \xi, \theta\rangle^2 \leq \|\theta\|^4\}.$$

Since $n\xi^\top \Sigma \xi =_d w^\top \Lambda^2 w = \|\Lambda w\|^2, \mathbb{E}(\xi^\top \Sigma \xi) = \operatorname{Tr}(\Sigma^2)/n$ and $\sqrt{n}\xi^\top \theta =_d \theta^\top V \Lambda^{1/2} w$ where w has independent $1-$sub-Gaussian entries. Using Lemma 11, we get that

$$\mathbb{P}(\mathscr{A}^c) \leq \delta + e^{-cn\|\theta\|^4/\theta^\top \Sigma \theta} \leq \delta + e^{-cn},$$

for some $c > 0$ small enough. Observe that, on event $\mathscr{A}$, we have

$$e^{-c\frac{\|\theta\|^4 - 2\langle \xi, \theta\rangle^2}{4\|\Sigma\|_\infty \|\theta\|^2 + 4\xi^\top \Sigma \xi}} \leq \exp\left(-c\frac{\|\theta\|^4}{\theta^\top \Sigma \theta + \frac{\operatorname{Tr}(\Sigma^2) + \|\Sigma\|_\infty^2 \log(1/\delta)}{n}}\right).$$

This show the result in probability. We next assume that $\|\theta\|^2 \geq \Delta^2 \|\Sigma\|_\infty$. Replacing $\log(1/\delta)$ by $cn\Delta^2$, we get that

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

Moreover, on event $\mathscr{A}$, we have now

$$e^{-c\frac{\|\theta\|^4 - 2\langle \xi, \theta\rangle^2}{4\|\Sigma\|_\infty \|\theta\|^2 + 4\xi^\top \Sigma \xi}} \leq \exp\left(-c\frac{\Delta^4}{\Delta^2 + \frac{r(\Sigma^2)}{n}}\right).$$

Therefore, we see that

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

We conclude that

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

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

A.3 Proof of Proposition 3.2

Let θ be a vector in $\mathbb{R}^p$. For the rest of the proof, we use the notation $\hat{\eta} := \hat{\eta}_{\mathrm{LDA}}$. We start by observing that

$$\mathbb{P}((\hat{\eta}(Y_{n+1}) \neq \eta_{n+1}) = \mathbb{P}\left(\left\langle \sum_{i=1}^n \Sigma^{-1/2} Y_i \eta_i, \Sigma^{-1/2} Y_{n+1} \eta_{n+1} \right\rangle < 0\right).$$

Let us denote $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \Sigma^{-1/2} Y_i \eta_i = \Sigma^{-1/2}\theta + \xi$ where $\xi = \frac{1}{n}\sum_{i=1}^n \Sigma^{-1/2} W_i \eta_i$. We get the following lower bound under the Gaussian noise:

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

for some $c, C > 0$, where we have conditioned on $\hat{\theta}$ to get the last inequality and used the fact that $\eta_{n+1}\Sigma^{-1/2}W_{n+1}$ has i.i.d standard Gaussian entries. Next, we have that

$$\langle \hat{\theta}, \Sigma^{-1/2}\theta\rangle^2 = \left\langle \Sigma^{-1/2}\theta + \xi, \Sigma^{-1/2}\theta \right\rangle^2 \leq 2\|\theta\|_\Sigma^4 + 2\left\langle \xi, \Sigma^{-1/2}\theta\right\rangle^2,$$

and that

$$\|\hat{\theta}\|^2 \geq \|\theta\|_\Sigma^2 + \|\xi\|^2 - 2\left|\theta^\top \Sigma^{-1/2}\xi\right|.$$

This implies that

$$\mathbf{E}(\mathscr{R}(\hat{\eta})) \geq C\mathbf{E}\left(e^{-c\frac{2\|\theta\|_\Sigma^4 + 2\left\langle \xi, \Sigma^{-1/2}\theta\right\rangle^2}{\|\theta\|_\Sigma^2 + \|\xi\|^2 - 2\left|\theta^\top\Sigma^{-1/2}\xi\right|}}\right).$$

Let us define the event

$$\mathscr{A} = \{\|\xi\|^2 \geq p/(2n)\} \cap \left\{\left|\theta^\top \Sigma^{-1/2}\xi\right| \leq \|\theta\|_\Sigma/8\right\}.$$

Since $n\|\xi\|^2 =_d \|w\|^2, \mathbb{E}(\|\xi\|^2) = p/n$ and $\sqrt{n}\xi^\top\Sigma^{-1/2}\theta =_d \|\theta\|_\Sigma w_1$, where $=_d$ denotes equality in distribution and $w = (w_1, \ldots, w_p)^T$ has independent standard Gaussian entries, it is easy to see that

$$\mathbb{P}(\mathscr{A}^c) \leq e^{-cp} + e^{-cn} \leq 1/2,$$

for some $c > 0$ small enough and n, p large enough. Observe that, on event $\mathscr{A}$, we have the inequality

$$e^{-c\frac{2\|\theta\|_\Sigma^4 + 2\left\langle \xi, \Sigma^{-1/2}\theta\right\rangle^2}{\|\theta\|_\Sigma^2 + \|\xi\|^2 - 2\left|\theta^\top\Sigma^{-1/2}\xi\right|}} \geq e^{-c\frac{2\|\theta\|_\Sigma^4 + 2\|\theta\|_\Sigma^2}{\|\theta\|_\Sigma^2 + p/(2n) - \|\theta\|_\Sigma/4}}.$$

Therefore,

$$\mathbf{E}(\mathscr{R}(\hat{\eta})) \geq C\mathbf{E}\left(e^{-c\frac{2\|\theta\|_\Sigma^4 + 2\left\langle \xi, \Sigma^{-1/2}\theta\right\rangle^2}{\|\theta\|_\Sigma^2 + \|\xi\|^2 - 2\left|\theta^\top\Sigma^{-1/2}\xi\right|}}\mathbf{1}\{\mathscr{A}\}\right) \geq C' e^{-c\frac{2\|\theta\|_\Sigma^4 + 2\|\theta\|_\Sigma^2}{\|\theta\|_\Sigma^2 + p/(2n) - \|\theta\|_\Sigma/4}}$$

for $C' = C/2$. If $\|\theta\|_\Sigma \geq 1/2$, the result is straightforward. Otherwise, if $\|\theta\|_\Sigma \leq 1/2$, then

$$\frac{2\|\theta\|_\Sigma^4 + 2\|\theta\|_\Sigma^2}{\|\theta\|_\Sigma^2 + p/(2n) - \|\theta\|_\Sigma/4} \leq \frac{C''n}{p} \leq C''',$$

and the bound is constant in this case. As a conclusion, we obtain the inequality

$$\mathbf{E}(\mathscr{R}(\hat{\eta}_{\mathrm{LDA}})) \geq C' \exp\left(-c\frac{\|\theta\|_\Sigma^4}{\|\theta\|_\Sigma^2 + \frac{p}{n}}\right)$$

for some absolute constants $c, C' > 0$.

A.4 Proof of Proposition 3.1

Let θ be a vector in $\mathbb{R}^p$. For the rest of the proof, we use the notation $\hat{\eta} := \hat{\eta}_{\text{ave}}$. We follow similar steps as in the proof of Proposition 3.2:

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

Let us denote $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i\eta_i = \theta + \xi$ where $\xi = \frac{1}{n}\sum_{i=1}^{n} W_i\eta_i$. We get the following lower bound under the Gaussian noise:

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

for some $c, C > 0$, where we have conditioned on $\hat{\theta}$ to get the last inequality and used the fact that $\eta_{n+1}\Sigma^{-1/2}W_{n+1}$ has i.i.d standard Gaussian entries. Next, we have that

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

and that

$$\hat{\theta}^\top\Sigma\hat{\theta} \geq \theta^\top\Sigma\theta + \xi^\top\Sigma\xi - 2\left|\theta^\top\Sigma\xi\right|.$$

Hence

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

Let us define the event

$$\mathscr{A} = \{\xi^\top\Sigma\xi \geq \mathrm{Tr}(\Sigma^2)/(2n)\} \cap \left\{\left|\theta^\top\xi\right| \leq \sqrt{\theta^\top\Sigma\theta}/8\right\} \cap \left\{\left|\theta^\top\Sigma\xi\right| \leq 4\sqrt{\theta^\top\Sigma^3\theta/n}\right\}.$$

Since $n\xi^\top\Sigma\xi =_d \|\Sigma w\|^2, \mathbb{E}(\xi^\top\Sigma\xi) = \mathrm{Tr}(\Sigma^2)/n$ and $\sqrt{n}\xi^\top v =_d v^\top\Sigma^{1/2}w$ for a given v, where $w = (w_1, \ldots, w_p)^T$ has independent standard Gaussian entries, it is easy to see, using the Hanson-Wright inequality Rudelson and Vershynin (2013a), that

$$\mathbb{P}(\mathscr{A}^c) \leq 1/4 + e^{-cn} + e^{-cr(\Sigma^2)} \leq 1/2$$

for some $c > 0$ small enough and $n, r(\Sigma^2)$ large enough. Observe that, on event $\mathscr{A}$, we have

$$e^{-c\frac{2\|\theta\|^4+2\langle\xi,\theta\rangle^2}{\theta^\top\Sigma\theta+\xi^\top\Sigma\xi-2|\theta^\top\Sigma\xi|}} \geq e^{-c\frac{2\|\theta\|^4+2\theta^\top\Sigma\theta}{\theta^\top\Sigma\theta+\mathrm{Tr}(\Sigma^2)/(2n)-8\|\Sigma\|_\infty\sqrt{\theta^\top\Sigma\theta/n}}},$$

so that

$$\mathbf{E}(\mathscr{R}(\hat{\eta})) \geq C\mathbf{E}\left(e^{-c\frac{2\|\theta\|^4+2\langle\xi,\theta\rangle^2}{\theta^\top\Sigma\theta+\xi^\top\Sigma\xi-2|\theta^\top\Sigma\xi|}}\right) \geq C'e^{-c\frac{2\|\theta\|^4+2\theta^\top\Sigma\theta}{\theta^\top\Sigma\theta+\mathrm{Tr}(\Sigma^2)/(2n)-8\|\Sigma\|_\infty\sqrt{\theta^\top\Sigma\theta/n}}}$$

for $C' = C/2$. Using the fact that

$$8\|\Sigma\|_\infty\sqrt{\theta^\top\Sigma\theta/n} \leq \theta^\top\Sigma\theta/2 + 32\|\Sigma^2\|_\infty/n,$$

and that $r(\Sigma^2)$ is large enough, we deduce that

$$\mathbf{E}(\mathscr{R}(\hat{\eta})) \geq C' e^{-c' \frac{2\|\theta\|^4 + 2\theta^\top \Sigma \theta}{\theta^\top \Sigma \theta/2 + \mathrm{Tr}(\Sigma^2)/(2n)}},$$

for some $c' > 0$. If $\theta^\top \Sigma \theta \leq \|\theta\|^4$ the result is straightforward. Otherwise, the bound is of order constant. As a conclusion we get

$$\mathbf{E}(\mathscr{R}(\hat{\eta})) \geq C' e^{-c' \frac{\|\theta\|^4}{\theta^\top \Sigma \theta/2 + \mathrm{Tr}(\Sigma^2)/(2n)}},$$

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

Appendix B. Proofs related to the regularization/interpolation results

B.1 Proof of Theorem 3

Recall that

$$\mathscr{R}(\hat{\eta}_\lambda) \leq e^{-\frac{\langle \theta, \hat{\theta}_\lambda \rangle^2}{2\hat{\theta}_\lambda^\top \Sigma \hat{\theta}_\lambda}}, \tag{9}$$

conditionally on $\hat{\theta}_\lambda$. Observe that $\hat{\theta}_\lambda = \theta x/n + W^\top A_\lambda^{-1}\eta/n$ where $A_\lambda = \lambda I_n + Y^\top Y/n$ and $x = \eta^\top A_\lambda^{-1}\eta$. The risk is invariant by rescaling $\hat{\theta}_\lambda$, hence we rescale it by n/x. Therefore, without loss of generality, we may assume that $\hat{\theta}_\lambda = \theta + W^\top H_\lambda^{-1}\eta/x$. Using Lemma 9, we deduce that

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

On the one hand,

$$|\langle \theta, \hat{\theta}_\lambda \rangle| \geq \|\theta\|^2 - \theta^\top W A_\lambda^{-1} W^\top \theta/n - \frac{|\eta^\top A_\lambda^{-1} W^\top \theta|}{\eta^\top A_\lambda^{-1}\eta} + \frac{\left(\theta^\top W A_\lambda^{-1}\eta\right)^2}{n \cdot \eta^\top A_\lambda^{-1}\eta}.$$

On the other hand,

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

We will now control the numerator and denominator in the exponent in the right-hand side of (9) separately.

- Control of the numerator:

  Let us denote $\theta_* := \pi_{k^*}\theta$ and $\bar{\theta} = \theta - \theta^*$. Observing that $I_p - W A_\lambda^{-1} W^\top/n$ is nonnegative definite with spectral norm less than 1, we have

  $$\theta^\top \left(I_p - W A_\lambda^{-1} W^\top/n\right)\theta \geq \bar{\theta}^\top \left(I_p - W A_\lambda^{-1} W^\top/n\right)\bar{\theta}/2 - \theta_*^\top \left(I_p - W A_\lambda^{-1} W^\top/n\right)\theta_*.$$

Therefore,

$$|\langle\theta,\hat\theta_\lambda\rangle| \geq \|\bar\theta\|^2/2 - \|A_\lambda^{-1/2}W^\top\bar\theta\|^2/(2n) - \|\theta_*\|^2 - \frac{\|A_\lambda^{-1/2}W^\top\theta\|}{\sqrt{\eta^\top A_\lambda^{-1}\eta}}.$$

Using Lemma 12 and Lemma 13, we get that

$$|\langle\theta,\hat\theta_\lambda\rangle| \geq \|\bar\theta\|^2/4 - C_1\frac{\bar\theta^\top\Sigma\bar\theta}{\sum_{i>k^*}\lambda_i/n+\lambda} - C_2\sqrt{\theta^\top\Sigma\theta},$$

as $\|\theta_*\|^2 \leq \|\bar\theta\|^2/4$. Since $\bar\theta^\top\Sigma\bar\theta \leq \lambda_{k^*+1}\|\bar\theta\|^2$,

$$|\langle\theta,\hat\theta_\lambda\rangle| \geq \|\bar\theta\|^2/8 - C_2\sqrt{\theta^\top\Sigma\theta} \geq \|\theta\|^2/10 - C_2\sqrt{\theta^\top\Sigma\theta}.$$

If $\|\theta\|^2/10 \leq 2C_2\sqrt{\theta^\top\Sigma\theta}$, then the bound in Theorem 3 is trivial. We conclude that

$$|\langle\theta,\hat\theta_\lambda\rangle| \geq C_3\|\theta\|^2,$$

for some $C_3 > 0$.

- Control of the denominator in (9):

$$\hat\theta_\lambda^\top\Sigma\hat\theta_\lambda \leq C\Bigg(\theta^\top\Sigma\theta + \|\Sigma^{1/2}WA_\lambda^{-1}W^\top\theta\|^2/n^2 + \frac{2+2(\eta^\top A_\lambda^{-1}W^\top\theta/n)^2}{(\eta^\top A_\lambda^{-1}\eta)^2}\eta^\top A_\lambda^{-1}W^\top\Sigma WA_\lambda^{-1}\eta\Bigg).$$

Using the same bound as for the numerator and the fact that

$$\|W^\top\theta\|^2 \leq Cn\theta^\top\Sigma\theta$$

with probability at least $1-e^{-cn}$, we conclude that

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

Employing Lemma 11, we see that with probability at least $1-\delta$

$$\eta^\top A_\lambda^{-1}W^\top\Sigma WA_\lambda^{-1}\eta \leq 3/2\,\mathrm{Tr}(A_\lambda^{-1}W^\top\Sigma WA_\lambda^{-1}) + \|A_\lambda^{-1}W^\top\Sigma WA_\lambda^{-1}\|_\infty\log(1/\delta).$$

Hence, using Lemma 14 and Lemma 15, we get further that with probability $\geq 1-\delta$

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

and that

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

Notice also that

$$\frac{\eta^\top A_\lambda^{-1}W^\top\Sigma W A_\lambda^{-1}\eta}{n^2} \le C_1 + C_2\,(k^*+1)\log(1/\delta)/n.$$

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

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

Taking $\delta_1 = e^{-cn}$ yields that

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

In the display above, we have used the fact that k is the smallest integer that satisfies $r_k(\Sigma)+\lambda n/\lambda_{k+1} > bn$ for $b \ge 1$. We treat two cases separately:
(a) If $\mathrm{Tr}(\Sigma)/n+\lambda \ge b\|\Sigma\|_\infty$ then we can take $k^*=0$.
(b) If $\mathrm{Tr}(\Sigma)/n+\lambda \le b\|\Sigma\|_\infty$, then $k^*\ge 1$ and $\sum_{i>k^*}\lambda_i/n+\lambda \le b\lambda_{k^*}$.

B.2 Proof of Proposition 4.1

Let $\theta \in \mathscr{C}_{k^*}(\Sigma)$. If $k^*=0$, the result follows immediately from Theorem 3. We assume that $k^*>0$ (hence, $\sum_{i>k^*}\lambda_i/n+\lambda \le b\|\Sigma\|_\infty$). Following the same steps as in the proof if Theorem 3, we have the bound for the numerator of the form

$$|\langle\theta,\hat\theta_\lambda\rangle| \ge C_3\|\theta\|^2,$$

for some $C_3>0$ and with probability $\ge 1-e^{-cn}$. As for the denominator,

$$\begin{aligned}\hat\theta_\lambda^\top\Sigma\hat\theta_\lambda \le 2\Bigg(&\left\|\Sigma^{1/2}\left(I_p - WA_\lambda^{-1}W^\top/n\right)\theta\right\|^2 \\ &+ \frac{2+2(\eta^\top A_\lambda^{-1}W^\top\theta/n)^2}{(\eta^\top A_\lambda^{-1}\eta)^2}\eta^\top A_\lambda^{-1}W^\top\Sigma W A_\lambda^{-1}\eta\Bigg) \\ &\le 2\|\theta\|^2\|\Sigma\|_\infty + C\Big((\sum_{i>k^*}\lambda_i/n+\lambda)^2 \\ &\quad + \left(\sum_{i>k^*}\lambda_i/n+\lambda\right)\|A_\lambda^{-1/2}W^\top\theta\|^2/n\Big)\frac{\eta^\top A_\lambda^{-1}W^\top\Sigma W A_\lambda^{-1}\eta}{n^2},\end{aligned}$$

again with probability $\geq 1 - e^{-cn}$. Since $\|A_\lambda^{-1/2} W^\top\|_\infty^2 / n \leq 1$, then

$$\hat{\theta}_\lambda^\top \Sigma \hat{\theta}_\lambda \leq \|\theta\|^2 \|\Sigma\|_\infty,$$

with probability $\geq 1 - e^{-cn}$. Hence,

$$\sup_{\substack{\|\theta\| \geq \Delta \|\Sigma\|_\infty \\ \theta \in \mathscr{C}_{k^*}(\Sigma)}} \mathbf{E}(\mathscr{R}(\hat{\eta}_\lambda)) \leq C \exp\left(-c \frac{\Delta^4}{\Delta^2 + \frac{r(\Sigma^2)}{n}} \right) + e^{-cn}.$$

B.3 Proof of Theorem 5

Arguing as in the proof of Lemma 9, we have that

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

where $A = W^\top W / n$, $u = \|\theta\| \eta / \sqrt{n}$ and $v = W^\top \theta / \sqrt{n \|\theta\|^2}$ and $det > 0$. Hence $e_1^\top (Y^\top Y)^{-1} \eta$ has the same 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 10, we also have that

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

Notice that $\pi = \tilde{W} (\tilde{W}^\top \tilde{W})^{-1} \tilde{W}^\top$ is a $p \times p$ projection matrix. Since W is of full rank with overwhelming probability, $\|W_1\|^2 - W_1^\top \pi W_1 = W_1^\top (I_p - \pi) W_1 > 0$ with high probability as well. 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 employ the bound

$$\eta^\top (W^\top W)^{-1} W^\top \theta \leq \sqrt{\eta A_0^{-1} \eta} \|A_0^{-1/2} W^\top \theta\| / n \leq C \frac{\sqrt{\theta^\top \Sigma \theta}}{\sum_{i > k^*} \lambda_i / n}$$

that holds with probability $\geq 1 - e^{-cn}$. Under the condition

$$\sqrt{\theta^\top \Sigma} \theta \leq 1/(2C) \sum_{i > k^*} \lambda_i / n,$$

we have that

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

Next, observe that $\eta_1 W_1^\top (I_p - \pi) \theta$ is 1-sub-Gaussian with parameter $\|\Sigma^{1/2} (I_p - \pi) \theta\|$. Using the bound from the proof of Theorem 3, we get that with probability at least $1 - e^{-cn}$,

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

The display above yields that with probability at least $1 - C/n^2$,

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

under the conditions of the theorem. Finally, we have that $\eta_1 W_1^\top \tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\eta}$ is $1-$sub-Gaussian with parameter $\|\Sigma^{1/2}\tilde{W}(\tilde{W}^\top\tilde{W})^{-1}\tilde{\eta}\|$. Following the steps of the proof of Theorem 4.1, we see that

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

Therefore,

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

Finally, with probability at least $1 - C/n^2$,

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

We conclude that the desired result is true under the conditions $k^*\log^2(n) \leq Cn$ and $\sum_{i>k^*}\lambda_i^2 n\log(n) \leq C(\sum_{i>k^*}\lambda_i)^2$ the result follows. We can now apply the union bound over $i = 1, \ldots, n$ and deduce that the final bound holds with probability at least $1 - 1/n$.

B.4 Proof of Theorem 6

We will follow the same steps as in the proof of Theorem 3. Recall that, under our contamination framework, the new observations $\tilde{Y}_i$ are such that

$$\theta\eta_i + \tilde{\Sigma}^{1/2}w_i,$$

where w_i are i.i.d. standard normal vectors and $\tilde{\Sigma} := \mathbf{I}_p + O$ where $O = \sum_{i\in R} O_i e_i e_i^\top$. Recall that

$$\mathscr{R}(\hat{\eta}_0) \leq e^{-\frac{\langle\theta,\hat{\theta}_0\rangle^2}{2\|\hat{\theta}_0\|^2}}, \tag{10}$$

conditionally on $\hat{\theta}_0$. We denote $W_i = \tilde{\Sigma}^{1/2}w_i$. On the one hand, we have

$$|\langle\theta,\hat{\theta}_0\rangle| \geq \|\theta\|^2 - \theta^\top W A_0^{-1}W^\top\theta/n - \frac{|\eta^\top A_0^{-1}W^\top\theta|}{\eta^\top A_0^{-1}\eta} + \frac{\left(\theta^\top W A_0^{-1}\eta\right)^2}{n\cdot\eta^\top A_0^{-1}\eta}.$$

On the other hand,

$$\|\hat{\theta}_0\|^2 \le 2\left(\|\theta\|^2 + \frac{2+2(\eta^\top A_0^{-1}W^\top\theta/n)^2}{(\eta^\top A_0^{-1}\eta)^2}\eta^\top A_0^{-1}W^\top W A_0^{-1}\eta\right).$$

We will now control the numerator and the denominator separately.

- Control of the numerator in (10):

  Let us denote $\theta_* := \pi_r\theta$ and $\bar{\theta} = \theta - \theta^*$. Observing that $I_p - WA_0^{-1}W^\top/n$ is positive semi-definite and has spectral norm at most 1, we deduce that

  $$\theta^\top\left(I_p - WA_0^{-1}W^\top/n\right)\theta \\ \ge \bar{\theta}^\top\left(I_p - WA_0^{-1}W^\top/n\right)\bar{\theta}/2 - \theta_*^\top\left(I_p - WA_0^{-1}W^\top/n\right)\theta_*.$$

  Hence,

  $$|\langle\theta,\hat{\theta}_0\rangle| \ge \|\bar{\theta}\|^2/2 - \|A_0^{-1/2}W^\top\bar{\theta}\|^2/(2n) - \|\theta_*\|^2 - \frac{\|A_0^{-1/2}W^\top\theta\|}{\sqrt{\eta^\top A_0^{-1}\eta}}.$$

  Using Lemma 12 and Lemma 13 and replacing k^* by r (this is possible since $r \le n/4$), we get that

  $$|\langle\theta,\hat{\theta}_0\rangle| \ge \|\bar{\theta}\|^2/4 - C_1\frac{\|\bar{\theta}\|^2}{(p-r)/n} - C_2\|\theta\|\sqrt{p/n},$$

  as $\|\theta_*\|^2 \le \|\bar{\theta}\|^2/4$. Since $\|\bar{\theta}\|^2 \le \|\bar{\theta}\|^2$,

  $$|\langle\theta,\hat{\theta}_0\rangle| \ge \|\bar{\theta}\|^2/8 - C_2\|\theta\| \ge \|\theta\|^2/10 - C_2\|\theta\|\sqrt{p/n}.$$

  Using the fact that $\|\theta\|^2 = \Omega(p/n)$, we conclude that

  $$|\langle\theta,\hat{\theta}_0\rangle| \ge C_3\|\theta\|^2,$$

  for some $C_3 > 0$.

- Control of the denominator in (10): note that

  $$\|\hat{\theta}_0\|^2 \le C\left(\|\theta\|^2 + \frac{2+2(\eta^\top A_0^{-1}W^\top\theta/n)^2}{(\eta^\top A_0^{-1}\eta)^2}\eta^\top A_0^{-1}W^\top W A_0^{-1}\eta\right).$$

  Using the same bound as for the numerator and the fact that

  $$\|A_0^{-1/2}W^\top\theta\|^2 \le n\|\theta\|^2,$$

  we get further that

  $$\|\hat{\theta}\|^2 \le C\left(\|\theta\|^2 + ((p/n)^2 + \|\theta\|^2 p/n)\frac{\eta^\top A_0^{-1}W^\top W A_0^{-1}\eta}{n^2}\right).$$

In view of Lemma 11, with probability at least $1-\delta$

$$\eta^\top A_0^{-1} W^\top W A_0^{-1}\eta \le 3/2n\,\mathrm{Tr}(A_0^{-1}) + n\|A_0^{-1}\|_\infty \log(1/\delta).$$

Hence, we deduce that with probability $\ge 1-e^{-cn}$

$$\eta^\top A_0^{-1} W^\top W A_0^{-1}\eta \le Cn^3/p.$$

We conclude that

$$\|\hat\theta_0\|^2 \le C(\|\theta\|^2 + p/n)$$

with probability at least $1-e^{-cn}$.

B.5 Proof of Proposition 6.1

Define O such that $O = \lambda\sum_{i=1}^r e_i e_i^\top$ where $(e_1,\dots,e_p)$ is the canonical Euclidean basis of $\mathbb{R}^p$. Under our contamination model, the new observations $\tilde Y_i$ have the same distribution as

$$\theta\eta_i + \tilde W_i,$$

where W_i are i.i.d. centered Gaussian vectors with covariance $\mathbf{I}_p + O$. Let θ be a vector in $\mathbb{R}^p$. For the rest of the proof, we use the notation $\hat\eta := \hat\eta_{\mathrm{LDA}} = \hat\eta_{\mathrm{ave}}$. Recall that

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

Let us denote $\hat\theta = \frac{1}{n}\sum_{i=1}^n \tilde Y_i\eta_i = \theta + \xi$ where $\xi = \frac{1}{n}\sum_{i=1}^n \tilde W_i\eta_i$. We get the following lower bound under the Gaussian noise assumption:

$$\mathbb{P}((\hat\eta(Y_{n+1}) \ne \eta_{n+1}) = \mathbb{P}(\langle\hat\theta, \theta + \eta_{n+1}W_{n+1}\rangle \le 0) \ge C\mathbf{E}\left(e^{-c\frac{\langle\theta,\hat\theta\rangle^2}{\|\hat\theta\|^2}}\right),$$

for some $c, C > 0$, where we have conditioned on $\hat\theta$ to get the last inequality and used the fact that $\eta_{n+1}W_{n+1}$ has i.i.d standard Gaussian entries. Next, we note that

$$\langle\hat\theta,\theta\rangle^2 = \langle\theta+\xi,\theta\rangle^2 \le 2\|\theta\|^4 + 2\langle\xi,\theta\rangle^2,$$

and that

$$\|\hat\theta\|^2 \ge \|\theta\|^2 + \|\xi\|^2 - 2\left|\theta^\top\xi\right|.$$

Therefore,

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

Next, let us define the event

$$\mathscr{A} = \{\|\xi\|^2 \ge \lambda/8\} \cap \left\{\left|\theta^\top\xi\right| \le 10\sqrt{\lambda/n}\|\theta\|\right\}.$$

It is easy to see that

$$\mathbb{P}(\mathscr{A}^c) \le 1/2.$$

Observe that, on the event $\mathscr{A}$ and for n large enough

$$e^{-c\frac{2\|\theta\|^4+2\langle\xi,\theta\rangle^2}{\|\theta\|^2+\|\xi\|^2-2|\theta^\top\xi|}} \geq e^{-c'\frac{\|\theta\|^4+\|\theta\|^2\lambda/n}{\|\theta\|^2+\lambda}}.$$

Hence, for $\lambda = n$ and $\|\theta\|^2 \leq \sqrt{n}$ we conclude that

$$\mathbf{E}(\mathscr{R}(\hat{\eta}_{\mathrm{LDA}})) \wedge \mathbf{E}(\mathscr{R}(\hat{\eta}_{\mathrm{ave}})) \geq C',$$

for some $C' > 0$.

Appendix C. Auxiliary results from linear algebra

Lemma 7 *Let $u, v \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ be a symmetric, invertible matrix. Then*

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

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 $C^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}$. Therefore,

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

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.

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

We conclude observing that

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

■

Lemma 8 *Let $u, v \in \mathbb{R}^n$, and let $A \in \mathbb{R}^{n\times n}$ be a symmetric invertible 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 7 we have

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

Hence

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

The second part of the proof is also a straightforward computation. ■

Lemma 9 *For any $\lambda \geq 0$, we have that*

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

where $A_\lambda = \lambda I_n + W^\top W/n$. We also have for $\lambda > 0$ that

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

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

Proof Recall that $\hat{\theta}_\lambda = \theta + WH_\lambda^{-1}\eta/x$ where $H_\lambda = \lambda I_n + Y^\top Y/n$ and $x = \eta^\top H_\lambda^{-1}\eta$. Denote $w := \theta/\|\theta\|$ and note 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 8 that

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

where $\det_A := (1 - v^\top A^{-1} v) u^\top A^{-1} u + (1 + u^\top A^{-1} v)^2$. Consequently,

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

We also see that

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

As a conclusion, we get that

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

On the other hand,

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

By choosing $u = \theta$, $v = W\eta/n$ and $B = \lambda I_p + W W^\top / n$, we get from Lemma 8 that

$$\begin{aligned} \hat{\theta}_\lambda &= (u u^\top + u v^\top + v u^\top + B)^{-1} (u + v) \\ &= \frac{1}{\det_B} \left(B^{-1} u (1 - v^\top B^{-1} v) + B^{-1} v (1 + u^\top A^{-1} v) \right), \end{aligned}$$

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

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

■

Lemma 10 *Assume that $p > n$. Let $W = (W_1 \ \tilde{W}) \in \mathbb{R}^{p \times n}$ be a full rank matrix and $\omega = (\omega_1, \ \tilde{\omega}) \in \mathbb{R}^n$. Then*

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

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

Proof The right-hand side is well-defined. Since W is of full rank, $\tilde{W}^\top \tilde{W}$ is invertible and $\|W_1\|^2 > W_1^\top \pi W_1$. We will use the Schur complement formula (that holds as long as all matrix inverses exist) to write that

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

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

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

Using Sherman-Morrison formula we deduce 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 $E = (\tilde{W}^\top \tilde{W})^{-1} (\|W_1\|^2 - W_1^\top \pi W_1) + (\tilde{W}^\top \tilde{W})^{-1} \tilde{W}^\top W_1 W_1^\top \tilde{W} (\tilde{W}^\top \tilde{W})^{-1}$, Therefore,

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

Moreover,

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

and

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

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 from probability theory

Let us write the eigenvalue decomposition of Σ as $\Sigma = \sum_{i=1}^p \lambda_i v_i v_i^\top$, and expand $W^\top W$ in the basis of eigenvectors of Σ. More specifically, let $z_i = W^\top v_i/\sqrt{\lambda_i}$. Then $(z_i)_{1\le i\le n}$ is a basis of $\mathbb{R}^{n\times n}$ if we assume W to be full rank and $W^\top W = \sum_i \lambda_i z_i z_i^\top$. With our notations $A_\lambda = \frac{1}{n}W^\top W + \lambda I_n = \sum_i \lambda_i' z_i z_i^\top + \lambda I_n$, where $\lambda_i' = \lambda_i/n$. Let us also define

$$A_{-i} = \sum_{j\neq i} \lambda_j' z_j z_j^\top + \lambda I_n, \quad A_k = \sum_{i>k} \lambda_i' z_i z_i^\top + \lambda I_n,$$

and similarly

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

We can apply Lemmas 9 and 10 from Bartlett et al. (2020) to A_{-i}^0 and A_k^0. Adding λ to both sides, we get that there exists a constant $c > 0$ such that with probability at least $1 - 2e^{-n/c}$,

- For any $k \ge 0$

$$\frac{1}{c}\sum_{i>k}\lambda_i' - c\lambda_{k+1}' n + \lambda \le \lambda_n(A_k) \le \lambda_1(A_k) \le c\left(\sum_{i>k}\lambda_i' + \lambda_{k+1}' n\right) + \lambda.$$

- $\forall i \ge 1$,

$$\lambda_{k+1}(A_{-i}) \le \lambda_{k+1}(A_\lambda) \le \lambda_1(A_k) \le c\left(\sum_{i>k}\lambda_i' + \lambda_{k+1}' n\right) + \lambda.$$

- $1 \le i \le k$,

$$\lambda_n(A_\lambda) \ge \lambda_n(A_{-i}) \ge \lambda_n(A_k) \ge \frac{1}{c}\sum_{i>k}\lambda_i' - c\lambda_{k+1}' n + \lambda.$$

Recall that the k-th effective rank is defined by $r_k(\Sigma) := \frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}}$. Under the condition $r_k(\Sigma) + \frac{n\lambda}{\lambda_{k+1}} \ge bn$, the inequalities stated above yield that

- for $k = k^*$

$$\frac{1}{c}\sum_{i>k}\lambda_i/n + \lambda \le \lambda_n(A_k) \le \lambda_1(A_k) \le c\sum_{i>k}\lambda_i/n + \lambda.$$

- $\forall i \ge 1$,

$$\lambda_{k+1}(A_{-i}) \le \lambda_{k+1}(A_\lambda) \le \lambda_1(A_k) \le c\sum_{i>k}\lambda_i/n + \lambda.$$

- $1 \le i \le k^*$,

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

Lemma 11 *Let $\xi \in \mathbb{R}^p$ be a random vector with 1-sub-Gaussian i.i.d entries. and let $B \in \mathbb{R}^{p\times p}$ be a nonnegative definite matrix. Then with probability at least $1-2\delta$,*

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

for some absolute constant $C > 0$.

Proof Notice that $\mathbf{E}\xi^\top B\xi = \mathrm{Tr}(B)$. Using Hanson-Wright inequality Rudelson and Vershynin (2013b), we deduce that for any $t > 0$ and for an absolute constant $c > 0$,

$$\mathbb{P}\left(|\xi^\top B\xi - \mathrm{Tr}(B)| > t\right) \le 2\exp\left[-c\min\left(\frac{t^2}{\|B\|_F^2}, \frac{t}{2\|B\|_\infty}\right)\right].$$

It follows that for $t \ge \mathrm{Tr}(B)/2$, we have the inequality $\frac{t}{2\|B\|_\infty} \le \frac{t^2}{\mathrm{Tr}(B)\|B\|_\infty} \le \frac{t^2}{\|B\|_F^2}$. Therefore, we get the desired conclusion with $C = 2/c$. ■

Lemma 12 *With probability at least $1-e^{-cn}$,*

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

for some $c, C > 0$.

Proof Note that

$$\|A_\lambda^{-1/2}W^\top\theta\|^2 = \theta^\top W A^{-1}W^\top\theta \le \|W^\top\theta\|_2^2\|A^{-1}\|_\infty.$$

Since $W^\top\theta$ has the same distribution as $\|\theta\|_\Sigma \cdot \xi$, where ξ is a random vector with 1-sub-Gaussian i.i.d entries, Lemma 11 implies that with probability at least $1-e^{-cn}$,

$$\|A_\lambda^{-1/2}W^\top\theta\|^2 \le C\theta^\top\Sigma\theta n\|A_\lambda^{-1}\|_\infty.$$

Hence, we deduce that

$$\|A_\lambda^{-1/2}W^\top\theta\|^2 \le C\frac{n\theta^\top\Sigma\theta}{\sum_{i>k^*}\lambda_i/n+\lambda},$$

since $1/\|A_\lambda^{-1}\|_\infty = \lambda_n(A_\lambda) \ge \sum_{i>k}\lambda_i/n+\lambda$ for $k = k^*$. ■

Lemma 13 *There exist $c, C_1, C_2 > 0$ such that with probability at least $1-e^{-cn}$ we have*

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

and

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

Proof The first inequality is straightforward observing that

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

For the lower bound, we will the sub-Gaussian property of η. To lower bound $\mathrm{Tr}(A_\lambda^{-1})$ observe that

$$\mathrm{Tr}(A_\lambda^{-1}) = \sum_{i=1}^n \lambda_i(A_\lambda)^{-1} \ge \sum_{i=k^*+1}^n \frac{1}{c\lambda_{k^*+1}r_{k^*}(\Sigma)/n + \lambda} \ge \frac{n/c}{\lambda_{k^*+1}r_{k^*}(\Sigma)/n + \lambda}.$$

We conclude using Lemma 11. This is possible by choosing $\delta = e^{-c'n}$ with $c' > 0$ small enough since $n\|A_\lambda^{-1}\|_\infty \le c''\,\mathrm{Tr}(A_\lambda^{-1})$ for some constant $c'' > 0$. ■

Lemma 14 *With probability at least $1 - e^{-cn}$ we have*

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

for some $c > 0$.

Proof Let us denote $C := A^{-1}W^\top \Sigma W A^{-1}$. Using the Sherman–Morrison formula for the inverse of rank-1 perturbations, we get that

$$\begin{aligned}\mathrm{Tr}(C) &= \mathrm{Tr}(A^{-1}W^\top \Sigma W A^{-1}) \\ &= \sum_{i=1}^n \lambda_i^2 z_i(\lambda_i' z_i z_i^\top + A_{-i})^{-2} z_i^\top \\ &= \sum_{i=1}^n \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i' z_i^\top A_{-i}^{-1} z_i)^2}.\end{aligned}$$

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

$$\mathrm{Tr}(C) = \sum_{i=1}^l \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i' z_i^\top A_{-i}^{-1} z_i)^2} + \sum_{i>l} \lambda_i^2 z_i^\top A_\lambda^{-2} z_i.$$

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 $i \le k^*$, $\lambda_n(A_{-i}) \ge \frac{1}{c}\sum_{i>k^*}\lambda_i/n + \lambda$. Hence

$$z_i^\top A_{-i}^{-2} z_i \le \frac{c^2\|z_i\|^2}{\left(\sum_{i>k^*}\lambda_i/n + \lambda\right)^2}.$$

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

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

Using the sub-Gaussian property of z_i, the condition $k^* \leq n/c$ and the independence with $\mathscr{L}_i$, we see that with probability $1 - e^{-cn}$, for all $i = 1, \ldots, n$

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

The first sum can be bounded by

$$\sum_{i=1}^{k^*} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i' z_i^\top A_{-i}^{-1} z_i)^2} \leq c^4 \frac{n^2 \sum_{i=1}^{k^*} \|z_i\|^2}{\|\Pi_{\mathscr{L}_i} z_i\|^4} \leq c' k^* n.$$

For the second sum, consider the same event where $\lambda_n(A_\lambda) \geq \lambda_{k+1} r_k(\Sigma)/(nc_1) + \lambda$. On this event,

$$\sum_{i>k^*} \lambda_i^2 z_i^\top A_\lambda^{-2} z_i \leq \frac{c_1^2 \sum_{i>k^*} \lambda_i^2 \|z_i\|^2}{(\sum_{i>k^*} \lambda_i/n + \lambda/c_1)^2} \leq \frac{c_5 n \sum_{i>k^*} \lambda_i^2}{(\sum_{i>k^*} \lambda_i/n + \lambda/c_1)^2}.$$

Therefore,

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

for $0 \leq k^* \leq n/c$ with probability at least $1 - e^{-cn}$.

■

Lemma 15 *With probability at least* $1 - e^{-cn}$,

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

for some absolute constant $c > 0$.

Proof For the first half (elements corresponding to $i \leq k^*$) we simply bound the spectral norm by the trace. For the second part, we rely on the estimate

$$\left\| \sum_{i>k^*} \lambda_i^2 A_\lambda^{-1} z_i z_i^\top A_\lambda^{-1} \right\|_\infty \leq \|A_\lambda^{-1}\|_\infty^2 \left\| \sum_{i>k^*} \lambda_i^2 z_i z_i^\top \right\|_\infty.$$

Using the previously employed arguments, we see that with probability at least $1 - e^{-cn}$

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

Hence,

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

We conclude that

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

■

References

N. Ardeshir, C. Sanford, and D. Hsu. Support vector machines and linear regression coincide with very high-dimensional features. 2021.

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

M. Belkin, D. Hsu, and P. Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. arXiv preprint arXiv:1806.05161, 2018.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 116(32):15849–15854, 2019a.

M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? In The 22nd International Conference on Artificial Intelligence and Statistics, pages 1611–1619. PMLR, 2019b.

P. J. Bickel, B. Li, A. B. Tsybakov, S. A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. van der Vaart. Regularization in statistics. Test, 15:271–344, 2006.

X. Bing and M. Wegkamp. Interpolating discriminant functions in high-dimensional gaussian latent mixtures. 2022. doi: 10.48550/ARXIV.2210.14347. URL `https://arxiv.org/abs/2210.14347`.

T. T. Cai and L. Zhang. High-dimensional linear discriminant analysis: optimality, adaptive algorithm, and missing data. 2018.

Y. Cao, Q. Gu, and M. Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. 2021.

X. Chen and A. Y. Zhang. Optimal clustering in anisotropic gaussian mixture models. arXiv preprint arXiv:2101.05402, 2021.

G. Chinot and M. Lerasle. On the robustness of the minimum ℓ_2 interpolator. 2021.

G. Chinot, F. Kuchelmeister, M. Löffler, and S. van de Geer. Adaboost and robust one-bit compressed sensing. 2021.

D. Davis, M. Diaz, and K. Wang. Clustering a mixture of gaussians with unknown covariance. 2021.

E. De Vito, L. Rosasco, and A. Rudi. Regularization: From inverse problems to large-scale machine learning. Harmonic and Applied Analysis: From Radon Transforms to Machine Learning, pages 245–296, 2021.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. Annals of statistics, 50(2):949, 2022.

D. Hsu, V. Muthukumar, and J. Xu. On the proliferation of support vectors in high dimensions. 2020.

T. Liang and A. Rakhlin. Just interpolate: kernel "ridgeless"' regression can generalize. The Annals of Statistics, 48(3):1329–1347, 2020.

T. Liang and B. Recht. Interpolating classifiers make few mistakes. 2021.

V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai. Classification vs regression in overparameterized regimes: does the loss function matter? 2020.

M. Ndaoud. Sharp optimal recovery in the two component gaussian mixture model. The Annals of Statistics, 50(4):2096–2126, 2022.

S. Oymak and M. Soltanolkotabi. Overparameterized nonlinear learning: gradient descent takes the shortest path? In Proceedings of the 36th International Conference on Machine Learning, pages 4951–4960, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/oymak19a.html`.

M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration, 2013a.

M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013b.

M. J. Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. Annual Review of Statistics and Its Application, 1:233–253, 2014.

K. Wang and C. Thrampoulidis. Benign overfitting in binary classification of gaussian mixtures. arXiv preprint arXiv:2011.09148, 2020.

K. Wang, Y. Yan, and M. Diaz. Efficient clustering for stretched mixtures: landscape and optimality. arXiv preprint arXiv:2003.09960, 2020.

D. Wu and J. Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. arXiv preprint arXiv:2006.05800, 2020.