Title

Benign Overfitting in Linear Regression and Classification

Author

Tsigler, Alexander

Publication Date

2024

Benign Overfitting in Linear Regression and Classification

by

Alexander Tsigler

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter L. Bartlett, Chair
Professor Adityanand Guntuboyina
Professor Anant Sahai

Summer 2024

Benign Overfitting in Linear Regression and Classification

Abstract

Benign Overfitting in Linear Regression and Classification

by

Alexander Tsigler

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Peter L. Bartlett, Chair

Benign overfitting, a phenomenon where deep neural networks predict well despite perfectly fitting noisy training data, challenges classical statistical intuition, which suggests a tradeoff between training data fit and prediction rule complexity. This dissertation explores benign overfitting in the context of linear models in the overparameterized regime, that is, where the dimension exceeds the number of data points. We study both regression and classification settings, focusing on the ridge regression solution, particularly its special case of zero regularization known as the minimum norm interpolating (MNI) solution.

In regression, we show that for MNI to exhibit benign overfitting, the data must possess a specific structure: data points should be nearly orthogonal when projected onto a subspace of small co-dimension. Learning occurs within the low-dimensional subspace, while the orthogonal complement absorbs noise, providing implicit regularization that adds to the explicit ridge regularization applied to the problem.

For classification, we study a scenario with two classes sharing the same covariance and opposite means, assuming the clusters exhibit the “benign structure” identified in regression. Our findings indicate that benign overfitting can also occur in classification, though the mechanism is more intricate. The ridge regression solution exhibits different regimes depending on the distance between the cluster centers.

To my family

Contents

List of Figures

List of Tables

Acknowledgments

I am deeply grateful to my advisor, Professor Peter L. Bartlett, whose profound expertise in deep learning theory and unwavering support made this work possible. His mentorship has been instrumental in my growth as a researcher.

I extend my gratitude to the Department of Statistics at UC Berkeley and especially to La Shana Porlaris, the Graduate Advisor and Director of Student Services, for guidance and assistance. The department's commitment to student success, clear program requirements, and financial support have been invaluable to my academic journey.

I am thankful to Professor Anant Sahai for the opportunity to serve as a Graduate Student Instructor for his Introduction to Machine Learning course. Being part of the content creation team and having my research incorporated into the curriculum significantly broadened my knowledge of machine learning and deepened my understanding of my own work through our insightful conversations.

Acknowledgments are due to the Simons Institute for the Theory of Computing, where I participated in workshops from the Foundations of Deep Learning program early in my doctoral research. These workshops shaped my understanding of the field, identifying open problems in deep learning, and clarifying the objectives of my own research projects.

The Geometric Functional Analysis and Applications program at MSRI (now SLMath) introduced me to Roman Vershynin's book on high-dimensional probability during my first year of PhD. The techniques and insights from this book have formed the foundation of my research, profoundly shaping the direction and methodology of my doctoral work.

Finally, I am grateful to my collaborators Philip M. Long, Gábor Lugosi, Luiz Chamon, and Spencer Frei for insightful discussions and suggestions.

Chapter 1

Introduction

1.1 Motivation

Deep learning methodology has revealed a surprising statistical phenomenon: interpolation can perform well. The classical perspective in statistical learning theory is that there should be a tradeoff between the fit to the training data and the complexity of the prediction rule. Whether complexity is measured in terms of the number of parameters, the number of non-zero parameters in a high-dimensional setting, the number of neighbors averaged in a nearest-neighbor estimator, the scale of an estimate in a reproducing kernel Hilbert space, or the bandwidth of a kernel smoother, this tradeoff has been ubiquitous in statistical learning theory.

Deep learning seems to operate outside the regime where results of this kind are informative. Deep neural networks can be overparameterized —having more trainable parameters than data points—and trained to perfectly fit (interpolate) the training data, but still generalize well to the test sample. We refer to this phenomenon as "benign overfitting."

As one example of benign overfitting, consider the experiment illustrated in Figure 1(c) in [61]: standard deep network architectures and stochastic gradient algorithms, run until they perfectly fit a standard image classification training set, give respectable prediction performance, *even when significant levels of label noise are introduced.* The deep networks in the experiments reported in [61] achieved essentially zero cross-entropy loss on the training data. In statistics and machine learning textbooks, an estimate that fits every training example perfectly is often presented as an illustration of overfitting ("... interpolating fits... [are] unlikely to predict future data well at all." [23, p37]).

Classical theory suggests that large models overfit the data and require significant regularization to generalize. In some cases, however, the best value of the regularizer can be zero [31] or even negative [28] for overparameterized models. Thus, to arrive at a scientific understanding of the success of deep learning methods, it is a central challenge to understand the performance of prediction rules that are trained with little or no regularization and can fit the training data perfectly.

In this dissertation, we consider perhaps the simplest class of models that can exhibit overparameterization and interpolation: linear models. The goal of this work is to mathematically characterize how benign overfitting can occur in such models and to provide high-level explanations of the mechanisms that allow for it.

1.2 Problem statement

We consider linear regression and classification settings, in which the covariates come as independent identically distributed (i.i.d.) samples from a distribution on $\mathbb{R}^p$, and the targets are real-valued. From this data, the learner infers a linear function on $\mathbb{R}^p$, which is then used for predicting the label of a newly sampled covariate vector.

For a linear model to be able to interpolate the data, the dimension p should be larger than the number of training data points n. Thus, we always assume that $p > n$.

Regression

In regression we assume that the covariate-target pairs $(\boldsymbol{x}_i, y_i)_{i=1}^n$ come as i.i.d. samples from a distribution on $\mathbb{R}^p \times \mathbb{R}$, and the responses are related to the covariates in the following way: y_i is generated as $y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}^* + \varepsilon_i$, where $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is the vector of coefficients of the ground truth linear model, and $(\varepsilon_i)_{i=1}^n$ are i.i.d. samples from a zero mean noise distribution with variance σ_ε^2. We denote the vector whose coordinates are $(y_i)_{i=1}^n$ as $\boldsymbol{y} \in \mathbb{R}^n$, and the matrix whose rows are $(\boldsymbol{x}_i^\top)_{i=1}^n$ as $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. That is, $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{Z}$ is a matrix with i.i.d. isotropic rows, and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{x}_i$.

We consider ridge regression solution, which for a regularization parameter $\lambda \in \mathbb{R}$ we define as follows:

$$\hat{\boldsymbol{\theta}} := \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_n)^{-1} \boldsymbol{y}.$$

If $(\boldsymbol{x}_{n+1}, y_{n+1})$ is a new independent sample of a covariate-target pair from the same distribution, then the excess risk of the prediction rule $\boldsymbol{x} \to \boldsymbol{x}^\top \hat{\boldsymbol{\theta}}$ is defined by the following formula:

$$\mathbb{E}_{\boldsymbol{x}_{n+1}, y_{n+1}} (\boldsymbol{x}_{n+1}^\top \hat{\boldsymbol{\theta}} - y_{n+1})^2 - \mathbb{E}_{\boldsymbol{x}_{n+1}, y_{n+1}} (\boldsymbol{x}_{n+1}^\top \boldsymbol{\theta}^* - y_{n+1})^2.$$

Here $\mathbb{E}_{\boldsymbol{x}_{n+1}, y_{n+1}}$ denotes expectation over the draw of $(\boldsymbol{x}_{n+1}, y_{n+1})$. The main technical goal of our analysis of regression is obtaining sharp non-asymptotic bounds on the excess risk.

An important particular case of the ridge regression solution arises when we set regularization λ to zero. In this case the solution becomes the minimum-norm interpolating solution (MNI), that is, the minimum norm vector $\boldsymbol{\theta}$ such that $\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$. Studying generalization of MNI is our main motivation, but extending the results to ridge regression allows us to make additional high-level conclusions. For example, we show that in an overparameterized setting the data itself can introduce “implicit regularization” that adds to the explicit regularization λ imposed on the problem. One unexpected qualitative conclusion is that somethimes this implicit regularization is too large, and choosing negative λ is optimal.

Classification

In classification we assume that first the true labels $(y_i)_{i=1}^n$ are sampled as i.i.d. Rademacher random variables (that is, $y_i = 1$ with probability 0.5, and $y_i = -1$ otherwise). Then the covariates are generated as $\boldsymbol{x}_i = \boldsymbol{q}_i + y_i\boldsymbol{\mu}$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the center of the positive cluster and $(\boldsymbol{q}_i)_{i=1}^n$ are i.i.d. samples from a centered distribution on $\mathbb{R}^p$. We denote the vector whose coordinates are $(y_i)_{i=1}^n$ as $\boldsymbol{y} \in \mathbb{R}^n$, and the matrix whose rows are $(\boldsymbol{x}_i^\top)_{i=1}^n$ as $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. That is, $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2} + \boldsymbol{y}\boldsymbol{\mu}^\top$, where $\boldsymbol{Z}$ is a matrix with i.i.d. isotropic rows, and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{q}_i$. The labels $\boldsymbol{y}$ are not actually known to the learner, and instead the learner observes the vector $\hat{\boldsymbol{y}}$, which is obtained from $\boldsymbol{y}$ by flipping the sign of each coordinate with some probability η.

As in regression, we consider the following ridge regression solution:

$$\hat{\boldsymbol{w}} := \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I}_n)^{-1}\hat{\boldsymbol{y}}.$$

If $(\boldsymbol{x}_{n+1}, y_{n+1})$ is a new independent sample from the same distribution, the misclassification probability of the linear classification rule $\boldsymbol{x} \to \operatorname{sign}(\boldsymbol{x}^\top\hat{\boldsymbol{w}})$, is defined as $\mathbb{P}(\operatorname{sign}(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{w}}) \neq y_{n+1})$. The main technical goal of our analysis of classification is obtaining sharp bounds for this misclassification probability.

Even though the regression and the classification setting that we consider look very similar, there is a fundamental difference. In regression the "signal vector" $\boldsymbol{\theta}^*$ belongs to the dual space: the matrix $\boldsymbol{X}$ does not depend on $\boldsymbol{\theta}^*$, and only the target vector $\boldsymbol{y}$ depends on its product with $\boldsymbol{X}$. In classification, however, the "signal vector" $\boldsymbol{\mu}$ belongs to the primal space: the vectors $(y_i\boldsymbol{\mu})_{i=1}^n$ are directly added to $(\boldsymbol{q}_i)_{i=1}^n$ to obtain $(\boldsymbol{x}_i)_{i=1}^n$. This distinction leads to significant differences in the analysis between regression and classification settings that we consider.

1.3 Covariance structure and technical assumptions

Let us fix the basis in the covariate space to be the eigenbasis of the covariance matrix $\boldsymbol{\Sigma}$. In this basis $\boldsymbol{\Sigma}$ is diagonal, that is,

$$\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_p).$$

Without loss of generality we can assume that the sequence $(\lambda_i)_{i=1}^p$ is non-increasing.

So far the elements of the sequence $(\lambda_i)_{i=1}^p$ are arbitrary parameters of the problem. The first result of this dissertation is that in the setting of regression, under the assumption that the data is Gaussian, a certain structure of this sequence is required for MNI to exhibit benign overfitting. More concretely, if all elements of $\boldsymbol{Z}$ are i.i.d. standard normal random variables, then for MNI to have small excess risk compared to the amount of noise σ_ε^2, there should exist k which is small compared to n, such that $\sum_{i>k}\lambda_i$ is large compared to $n\lambda_{k+1}$.

Introduce the following notation:

$$r_k := \frac{1}{\lambda_{k+1}}\sum_{i>k}\lambda_i.$$

The quantity r_0 is an important complexity parameter for covariance estimation problems, where it has been called the "effective rank" [56, 30]. Earlier, effective rank of $\boldsymbol{\Sigma}^2$ was called the "stable rank" [48] and the 'numerical rank" [49], although that term has a different meaning in computational linear algebra [20, p261]. A straightforward interpretation of effective rank is the effective number of dimensions across which the covariates are distributed: indeed, we are dividing the energy of the whole covariate vector $\sum_i \lambda_i$ by the maximum energy in a single direction λ_1. Thus, the condition for benign overfitting is that after removing the first k coordinates, the distribution of the covariates should be smeared across many more directions than the sample size. Throughout the dissertation we refer to such structure (or its variations) as "benign structure". The variations of the benign structure arise due to differences in technical assumptions imposed on the data and due to accounting for ridge regularization. The main idea, however, remains the same: after throwing out the first k coordinates, the covariates should have high effective dimension in some sense.

The separation of the first k eigendirections of $\boldsymbol{\Sigma}$ is in the core of all our results. Because of it, throughout this whole dissertation we use the following notation: for any $k \in \{0, 1, \dots, p\}$ and any matrix $\boldsymbol{M} \in \mathbb{R}^{n\times p}$ we denote $\boldsymbol{M}_{0:k}$ to be the matrix comprised of the first k columns of $\boldsymbol{M}$.[1] Analogously, we denote $\boldsymbol{M}_{k:\infty}$ to be the matrix comprised of the last $p-k$ columns of $\boldsymbol{M}$. For any $\boldsymbol{u} \in \mathbb{R}^p$ we denote $\boldsymbol{u}_{0:k}$ to be the vector comprised of the first k components of $\boldsymbol{u}$, and $\boldsymbol{u}_{k:\infty}$ — of the remaining components. Finally, we denote $\boldsymbol{\Sigma}_{0:k} = \operatorname{diag}(\lambda_1, \dots, \lambda_k)$ and $\boldsymbol{\Sigma}_{k:\infty} = \operatorname{diag}(\lambda_{k+1}, \dots, \lambda_p)$. We sometimes refer to the first k components as the "spiked part of the covariance"[2], and to the remaining components as the "tail of the covariance".

We choose the $k:\infty$ notation instead of $k:p$ to emphasize that our results don't depend on p, and only the notions of effective dimension implicitly given by the sequence $\{\lambda_i\}_{i=1}^p$ matter. For example, if one increases the dimension to $p' > p$ and pads the sequence $\{\lambda_i\}_{i=1}^p$ with $p'-p$ zeros, our results will still hold.

The assumption of Gaussianity that we mentioned above is not necessary. The first (and the most straightforward) generalization of this result is to the case when the elements of $\boldsymbol{Z}$ are independent and sub-Gaussian, as given by the following definition.

Definition 1. *For any centered random variable v we define its sub-Gaussian norm as*

$$\|v\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}\exp(v^2/t^2) \le 2\right\}.$$

If $\|v\|_{\psi_2} \le \sigma$, we say that the distribution of v is σ-sub-Gaussian.

Those assumptions, however, can also be weakened. The main point where our argument uses the independence of the coordinates is the proof that the matrix $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ has bounded condition number with high probability. A geometric interpretation of this fact is that

[1] When $k = 0$ this matrix is just empty and all the terms that involve $0:k$ index become zero.

[2] Here we use the word "spiked" as in the "spiked covariance models", which usually assume that the eigenvalues of $\boldsymbol{\Sigma}_{k:\infty}$ are all equal and of smaller order than eigenvalues of $\boldsymbol{\Sigma}_{0:k}$. One way to interpret our results is that only spiked-covariance-like models can exhibit benign overfitting, and we derive general conditions for a model to be spiked-covariance-like.

since the data has high effective rank in components $k:\infty$, the rows of $\boldsymbol{X}_{k:\infty}$ are almost orthogonal to each other, and their Gram matrix $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ behaves as a scaled identity matrix. Because of that, for our main bound on the excess risk in regression we make a direct assumption on the condition number of $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ instead of assuming that the components of $\boldsymbol{Z}$ are independent. We do still assume the rows of $\boldsymbol{Z}$ are sub-Gaussian, as given by the following definition.

Definition 2. *For any random vector $\boldsymbol{v}$ in $\mathbb{R}^p$ we define its sub-Gaussian norm as*

$$\|\boldsymbol{v}\|_{\psi_2} := \sup_{\boldsymbol{u}\in\mathbb{R}^p:\|\boldsymbol{u}\|=1} \|\boldsymbol{u}^\top \boldsymbol{v}\|_{\psi_2}.$$

If $\|\boldsymbol{v}\|_{\psi_2} \le \sigma$, we say that the distribution of $\boldsymbol{v}$ is σ-sub-Gaussian.

When it comes to classification, we assume that $\boldsymbol{\Sigma}$ has the benign structure described above from the very beginning. Similarly to regression, we make a direct assumption on the eigenvalues of $\boldsymbol{Z}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Z}_{k:\infty}$. Instead of assuming sub-Gaussianity, however, we precisely describe the event that should happen with high probability in order for our bound to hold. We prove that that event does hold with high probability under sub-Gaussianity, but argue that sub-Gaussianity can be relaxed.

1.4 Overview of the next chapters

We study regression in Chapter 2 and classification in Chapter 3.

Chapter 2 starts with deriving the "benign structure" of the covariance, that we introdced in Section 1.3. We discuss two simple settings first: "essentially high dimensional" and "essentially low dimensional". Then we derive general bounds for the variance term of the excess risk for MNI, which show that the covariance should decompose into an essentially low-dimensional and an essentially high-dimensional part in order for that term to be small. Given that decomposition, we proceed with deriving sharp bounds on the full excess risk of the ridge regression solution under weaker assumptions (as was discussed in Section 1.3). We show that the learning happens in the first k components, while the rest of the components absorb the noise and provide implicit regularization to the learning problem in the first k components. We then study the effect of ridge regularization on the bounds. One interesting effect happens when the implicit regularization coming from the data is very high: in this case it can be optimal to use a negative value of the ridge regularization parameter. Chapter 2 is based on the following publications:

1. Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* (2020). ISSN: 0027-8424. DOI: `10.1073/pnas.1907378117`. eprint: `https://www.pnas.org/content/early/2020/04/22/1907378117.full.pdf`. URL: `https://www.pnas.org/content/early/2020/04/22/1907378117`

2. Alexander Tsigler and Peter L. Bartlett. “Benign overfitting in ridge regression”. In: *Journal of Machine Learning Research* 24.123 (2023), pp. 1–76. URL: http://jmlr.org/papers/v24/22-1398.html

In Chapter 3 we also start with a discussion of the MNI solution, but with the purpose of investigating its geometric structure. Then, using the machinery that we developed in Chapter 2, and assuming that the distribution within the clusters has benign structure, we provide bounds on the classification accuracy. Furthermore, we study the bounds and show that the effect of the benign structure of the covariance is different than in regression. When label-flipping noise is not introduced (that is, $\eta = 0$), in components $k : \infty$ the ridge regression solution is approximately collinear with $\boldsymbol{\mu}_{k:\infty}$, but in components $0 : k$ it approximately recovers the optimal rotation of $\boldsymbol{\mu}_{0:k}$. That is, learning happens in both the spiked part of the covariance and in the tail, but it is somewhat more efficient in the spiked part. Introduction of the label-flipping noise (that is, setting η to be a small constant) does not qualitatively change the performance of the ridge regression solution if the magnitude of $\boldsymbol{\mu}$ is moderate. However, if $\boldsymbol{\mu}$ is large in magnitude, the performance of the ridge regression solution may change significantly, and the mechanism of benign overfitting in that regime becomes very similar to that in regression. Finally, we study the effect of ridge regularization in the setting without label-flipping noise. We show that one cannot achieve a significant gain in accuracy by increasing regularization beyond the point where the benign structure[3] appears. For example, just as in regression, if the implicit regularization coming from the data is large, it may be optimal to set ridge regularization parameter to a negative value. Chapter 3 is based on forthcoming work by Alexander Tsigler, Luiz Chamon, Spencer Frei and Peter L. Bartlett, expected to be posted in August 2024.

1.5 Additional notation

We use the symbol $:=$ to introduce definitions: for example, $b := a + 1$ means that we introduce a new quantity b which is defined as $a + 1$. We use $a \approx b$ to denote an informal statement that a and b are within a constant factor of each other with high probability (which we abbreviate as w.h.p.). Analogously, we use $a \gtrsim b$ ($a \lesssim b$) to denote “w.h.p. a is at least (at most) constant times b ”, and $a \gg b$ ($a \ll b$) to denote “w.h.p. a is much larger (smaller) than b”.

For any positive integer d we denote $\mathbf{0}_d \in \mathbb{R}^d$ to be the vector of all zeros, $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ to be the identity matrix. We use $\operatorname{diag}(a_1, \dots, a_d)$ to denote the diagonal matrix in $\mathbb{R}^{d\times d}$ whose diagonal elements are $a_1, \dots, a_d$.

We use $\mu_i(\boldsymbol{M})$ to denote the i-th largest eigenvalue of a symmetric matrix $\boldsymbol{M}$. For any square matrix $\boldsymbol{M}$ we denote its spectral norm by $\|\boldsymbol{M}\|$, its Frobenius norm by $\|\boldsymbol{M}\|_F$ and its trace by $\operatorname{tr}(\boldsymbol{M})$. For any $\boldsymbol{u} \in \mathbb{R}^d$ we denote its Euclidean norm by $\|\boldsymbol{u}\|$.

[3] Here we refer to a modified definition of benign structure, which takes regularization into account.

We abbreviate positive-definite as PD and positive-semi-definite as PSD. For any PSD matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ and any $\boldsymbol{u} \in \mathbb{R}^d$ we denote $\|\boldsymbol{u}\|_{\boldsymbol{M}} := \sqrt{\boldsymbol{u}^\top \boldsymbol{M} \boldsymbol{u}}$. For any matrix $\boldsymbol{M} \in \mathbb{R}^{n \times p}$ we denote its pseudo-inverse as $\boldsymbol{M}^\dagger \in \mathbb{R}^{p \times n}$.

We use $\mathbb{P}(\mathscr{A})$ to denote the probability of an event $\mathscr{A}$ and $\mathbb{E}[\xi]$ to denote expectation of a random variable ξ. We use the notation $\mathbb{P}_\xi$ and $\mathbb{E}_\xi$ for probability and expectation with respect to a draw of the random element ξ. We say that a random vector $\boldsymbol{v} \in \mathbb{R}^d$ is isotropic if $\mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top] = \boldsymbol{I}_d$. For a vector $\boldsymbol{m} \in \mathbb{R}^d$ and a PSD matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ we denote the normal distribution with mean $\boldsymbol{m}$ and covariance $\boldsymbol{M}$ as $\mathcal{N}(\boldsymbol{m}, \boldsymbol{M})$. If a random vector $\boldsymbol{v}$ has that distribution, we denote it as $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{M})$.

Chapter 2

Regression

2.1 Introduction

The aim of this chapter is to provide a theoretical understanding of benign overfitting in the setting of linear regression. As introduced in Section 1.2, our main goal is to bound and to analyze the excess risk of the ridge regression solution to an overparameterized linear regression problem.

Despite being a classical statistical methodology, ridge regression and its ridgeless limit were not completely studied by classical theory in such a regime: when $n < p$ it suggests that the regularization parameter should be large enough to provide additional capacity control (see, e.g., [25] and references therein). First, we study the variance term for ridgeless regression (that is, MNI) with $n < p$ under the additional assumption that the data vectors have independent components. We discover that the variance term can be small if and only if there exists $k \ll n$ such that if one removes the first k largest eigenvalues of the covariance operator, the remaining tail of the sequence of eigenvalues has large effective rank compared to n. After that, we start afresh and use the same separation of eigendirections from the very beginning, which allows us to substitute the independence assumption by a weaker assumption on the condition number of the Gram matrix of the tails of the data vectors. Moreover, we show how the same separation of the eigenvalues gives tight bounds for the bias term too. Finally, by virtue of algebra, our argument extends very easily to the setting of ridge regression, which allows for comparison with the above mentioned classical results and investigation of the case when the regularization is even less than zero. We show that we extend (with different constants) the results of [25] to a larger range of regularization parameters, and give general conditions under which negative regularization is optimal and can provide arbitrarily high multiplicative gain in excess risk.

The structure of the chapter is the following. We start by introducing the setup of ridge regression in Section 2.2, where we also derive the decomposition of the excess risk into the bias and variance terms. Then, in Section 2.3 we derive the necessity of the existence of k and provide an intuitive explanation of how its existence helps MNI interpolate the data without overfitting. After that we present our main bounds on the excess risk of the ridge regression solution in Section 2.4. Section 2.5 provides a technical discussion of the assumptions imposed in Section 2.4, and Section 2.6 provides an outline of the proof and explains where it uses the assumption that the data is sub-Gaussian. In Section 2.7 we note that as a side product of the proof an alternative form of the main bound arises, which makes it convenient to compare our bounds to other results. In Section 2.8, we derive the sufficient conditions for optimality of negative regularization. In Section 2.9, we provide an overview of the field of overparameterized ridge regression and explain how our work relates to others. Finally, we conclude the chapter with Section 2.10.

2.2 Ridge regression setup

The learning problem we consider is ridge regression. Its goal is to learn an unknown real-valued function on $\mathbb{R}^p$ given noisy observations of its values in n points. We operate in the overparameterized regime, i.e., $p > n$.

Covariate model

We assume that the data set consists of n i.i.d. vectors sampled from some distribution on $\mathbb{R}^p$, whose mean is zero. Throughout this chapter $\boldsymbol{x}$ denotes an independent draw from that distribution. Denote $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ to be the matrix whose rows are the (transposed) data vectors.

Our results depend on the spectrum of the covariance matrix $\boldsymbol{\Sigma}$. We fix an orthonormal basis in which $\boldsymbol{\Sigma}$ is diagonal:

$$\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_p), \tag{2.1}$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ is the non-increasing sequence of eigenvalues of $\boldsymbol{\Sigma}$.

We assume sub-Gaussianity: denote $\boldsymbol{Z} := \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ (whitened data matrix). Rows of $\boldsymbol{Z}$ are isotropic centered i.i.d. random vectors. We assume that rows of $\boldsymbol{Z}$ are σ_x-sub-Gaussian as defined in Section 1.3.

Sub-Gaussianity is a classical assumption, which provides a convenient framework for controlling deviations of various quantities of interest (see [55] for an introduction). We discuss whether it is actually needed in Section 2.6.

Response model

Denote $\boldsymbol{y} \in \mathbb{R}^n$ to be the vector whose coordinates are noisy measurements of the values of an unknown function in the corresponding data points. We assume that the true function is linear with coefficients $\boldsymbol{\theta}^* \in \mathbb{R}^p$, i.e.,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the noise vector. We assume that components of $\boldsymbol{\varepsilon}$ are i.i.d. centered random variables with variance $v_{\boldsymbol{\varepsilon}}^2$, and that $\boldsymbol{\varepsilon}$ is independent from $\boldsymbol{X}$.

Learning procedure

Ridge regression with regularization parameter λ is a classical learning algorithm that estimates $\boldsymbol{\theta}^*$ from $\boldsymbol{X}, \boldsymbol{y}$ according to the following formula:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y}) := \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_n)^{-1} y.$$

See Appendix A.1 for a discussion. The matrix $\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top$ will play an important role in our analysis, so we denote

$$\boldsymbol{A} := \lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top.$$

In the ridgeless case ($\lambda = 0$), $\boldsymbol{A}$ is the Gram matrix of the data. Ridge regularization shifts all its eigenvalues by λ.

Excess risk and its bias-variance decomposition

The quantity of interest is excess risk that we define in the following way: recall that $\boldsymbol{x}$ is a new data point from the same distribution as rows of $\boldsymbol{X}$. The error that our predictor incurs on this data point is $\boldsymbol{x}^\top(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}^*)$. We define excess risk as the average squared error over the population, i.e.,

$$\mathbb{E}_{\boldsymbol{x}}\left[(\boldsymbol{x}^\top(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}^*))^2\right] = \|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2.$$

Note that $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ is linear in $\boldsymbol{y}$, which allows us to write

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) + \hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}),$$
$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2\right] = \|\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 + \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2\right],$$
$$\|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 \le 2(\|\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 + \|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2).$$

The term $\|\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2$ is the error in the noiseless regime; it is caused by rows of $\boldsymbol{X}$ not spanning the whole space and by regularization. The term $\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2$ is the error of learning the zero function from pure noise. One can see that these two terms nicely decouple from each other and can be studied separately. Moreover, note that $\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2$ is a quadratic form in $\boldsymbol{\varepsilon}$. Its expectation scales linearly with $v_{\boldsymbol{\varepsilon}}^2$ (variance of the noise):

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2\right] = v_{\boldsymbol{\varepsilon}}^2 \mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{A}^{-1}).$$

If the noise is $\sigma_{\boldsymbol{\varepsilon}}$-sub-Gaussian, then by Lemma 82 from the appendix for some absolute constant c and any $t > 1$, with probability at least $1 - ce^{-n/c}$,

$$\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2 = \boldsymbol{\varepsilon}^\top\boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{A}^{-1}\boldsymbol{\varepsilon} \le ct\sigma_{\boldsymbol{\varepsilon}}^2 \mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{A}^{-1}).$$

Therefore, both expectation and deviations of the term $\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2$ are controlled by the quantity $\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{A}^{-1})$. Thus, we define:

$$\begin{aligned} B &:= \|\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 &&= \|(\boldsymbol{X}^\top\boldsymbol{A}^{-1}\boldsymbol{X} - \boldsymbol{I}_p)\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 && \text{— bias,} \\ V &:= \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2/v_{\boldsymbol{\varepsilon}}^2\right] &&= \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{A}^{-1}) && \text{— variance.} \end{aligned} \tag{2.2}$$

These quantities don't depend on the distribution of the noise. The goal of this chapter is to provide sharp non-asymptotic bounds for them.

2.3 Deriving the split into the spiked part and the tail

In this section we start by considering simple settings for which B and V are rather straightforward to assess. We then do a more involved computation for the term V to derive the benign covariance structure that we introduced in Section 1.3. The simple settings that we start with can be seen as building blocks, combining which gives that benign structure.

Essentially high-dimensional linear regression vs. essentially low-dimensional

Let us develop some intuition by considering two easy scenarios: "essentially low-dimensional" and "essentially high-dimensional". For each scenario we will do an informal computation of the excess risk and give a geometric interpretation.

- **Essentially low-dimensional linear regression.** Consider least squares regression in which data lives in $\mathbb{R}^k$ and $k \ll n$: $\boldsymbol{X} \in \mathbb{R}^{n\times k}$ with i.i.d. centered rows from a distribution with covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ has i.i.d. centered components with variances $v_{\boldsymbol{\varepsilon}}^2$. Our estimator of choice in this regime is OLS:

  $$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2 = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \boldsymbol{\varepsilon}\|^2.$$

  As $\boldsymbol{\theta}$ takes all possible values in $\mathbb{R}^k$, $\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ takes all possible values in the span of columns of $\boldsymbol{X}$, which means that

  $$\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \Pi_{\boldsymbol{X}}\boldsymbol{\varepsilon},$$

  where $\Pi_{\boldsymbol{X}}$ is the projection on the span of columns of $\boldsymbol{X}$. This allows us to write the following informal computation, which leads to the classical k/n rate:

  $$v_{\boldsymbol{\varepsilon}}^2 k = \mathbb{E}_{\boldsymbol{\varepsilon}}\|\Pi_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|^2 = \mathbb{E}_{\boldsymbol{\varepsilon}}\|\boldsymbol{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \underbrace{\boldsymbol{X}^\top\boldsymbol{X}}_{\approx n\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right],$$

  $$v_{\boldsymbol{\varepsilon}}^2 \cdot k/n \approx (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\langle \boldsymbol{x}, \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\rangle^2.$$

Here we used the informal transition $\|n^{-1}\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{\Sigma}\| \approx 0$ — the population covariance matrix is well-approximated by the sample covariance matrix uniformly in all directions. If $k \ll n$ this results holds with very few additional assumptions (see [52] and references therein).

What we have obtained is an example of a classical argument: the training error $\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|^2$ is a good proxy for the population error $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|^2$ uniformly over all $\hat{\boldsymbol{\theta}} \in \mathbb{R}^k$, and the model helps eliminate the noise because it gets projected on a subspace of low dimension. The larger the model, the more error comes from the noise.

Such a result leads to a classical bias-variance trade-off: the larger the model is, the better it can approximate the true dependence, but also the more noise it picks up. A classical cartoon is shown in Figure 2.1: Figures 2.1b–2.1d show the result of performing least squares regression with features $\{\cos(m\pi x)\}_{m=0}^p$. As the number of features grows, the ability of the model to approximate the signal grows too, but at the cost of increasing sensitivity to the noise. As the number of features approaches the number of data points (the "interpolation threshold"), this leads to overfitting.

- **Essentially high-dimensional linear regression.** Now consider linear regression in which $p \gg n$, but with isotropic data: assume that the matrix $\boldsymbol{X}$ has i.i.d. standard normal elements and $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_* + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}_n, v_{\boldsymbol{\varepsilon}}^2\boldsymbol{I}_n)$ — independent from $\boldsymbol{X}$. We consider the minimum ℓ_2-norm interpolating solution:

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p:\boldsymbol{X}\boldsymbol{\theta}=\boldsymbol{y}} \|\boldsymbol{\theta}\| = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}(\boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}).$$

According to our definitions of bias and variance from Equation (2.2) with $\lambda = 0$,

$$\begin{aligned} B =& \|\big(\boldsymbol{I}_p - \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\big)\boldsymbol{\theta}^*\|, \\ V =& \mathbb{E}_{\boldsymbol{\varepsilon}}\|\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\varepsilon}\|^2/v_{\boldsymbol{\varepsilon}}^2 = \operatorname{tr}\Big((\underbrace{\boldsymbol{X}\boldsymbol{X}^\top}_{\approx p\boldsymbol{I}_n})^{-1}\Big). \end{aligned}$$

Here we see the following: the matrix $\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}$ is the projection on the span of the data. This is a random n-dimensional subspace in p-dimensional space. Thus, with high probability $\|\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{\theta}^*\|^2/\|\boldsymbol{\theta}^*\|^2 \approx n/p$, so the projection only preserves an n/p fraction of the energy of the signal. When it comes to the variance term, we can use the same concentration result for the sample covariance as we did in the low-dimensional case, but for the transposed data matrix, meaning $\boldsymbol{X}\boldsymbol{X}^\top \approx p\boldsymbol{I}_n$. Finishing the computation yields

$$B \approx (1-n/p)\|\boldsymbol{\theta}^*\|^2, \quad \mathbb{E}_{\boldsymbol{\varepsilon}}V \approx n/p.$$

We see that the signal is almost not learned at all in this regime (the bias term is close to the full energy of the signal), but the noise is also damped by the factor p/n.

The geometric interpretation is as follows: if $p \gg n$, the span of n data points is almost orthogonal to $\boldsymbol{\theta}^*$ with high probability. The data just does not measure $\boldsymbol{\theta}^*$ in most directions, so almost the whole signal is lost. On the other hand, despite the noise fully propagating into in-sample predictions, a new data point $\boldsymbol{x}$ is also almost orthogonal to all the old data points with high probability, so those noisy predictions don't influence the prediction in $\boldsymbol{x}$. Overall, despite interpolating the data, we effectively learn a zero estimate out of sample. The zero estimator can be a very good estimator, e.g., if the true signal is zero. This hints at the possibility of learning via high-dimensional interpolation: the model can use the directions in which the signal is not learned to smear the noise over them.

The learning cartoon for this regime is given in Figures 2.2b–2.2c: as the number of cosine features becomes large compared to the number of data points, the learning procedure predicts zero out of sample, despite interpolating the values in sample. However, if we add certain multiplicative weights to the cosine features, down-weighting higher frequencies, it causes the minimum norm solution to learn the low frequency signal and interpolate the noise using the high frequency components.

Deriving k

In this section we explain how to derive the structure we introduced in Section 1.3. We only consider the MNI solution here (that is, we set λ to zero until the end of this section), and we impose an additional assumption that all the elements of the matrix $\boldsymbol{Z}$ are independent.

Recall that the variance term is defined as follows:

$$V = \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{A}^{-1}).$$

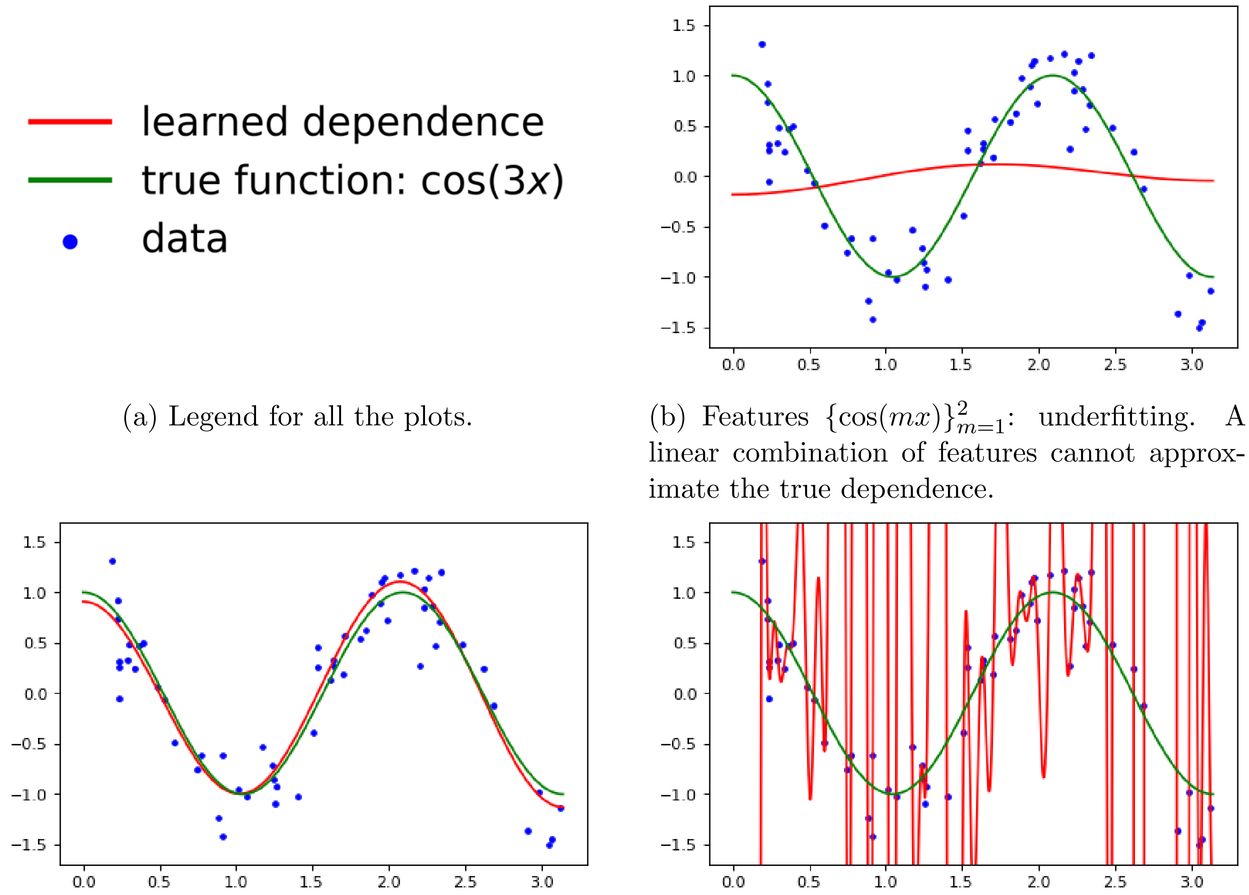
The first idea is to represent this quantity as a sum over the columns of matrix $\boldsymbol{X}$. Denote the i-th column of matrix $\boldsymbol{Z}$ as $\boldsymbol{z}_i$. Then we can write the following.

$$V = \sum_{i=1}^{p} \lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}^{-2}\boldsymbol{z}_i.$$

Recall that vectors $\boldsymbol{z}_i$ are random with i.i.d. components (as the rows of $\boldsymbol{X}$ are i.i.d.). That would allow the quantity $\boldsymbol{z}_i^\top\boldsymbol{A}^{-2}\boldsymbol{z}_i$ to be bounded by standard results on concentration of quadratic forms if the matrix $\boldsymbol{A}$ was independent from $\boldsymbol{z}_i$. This idea leads to the following. First, we write

$$\boldsymbol{A} = \sum_{j=1}^{p} \lambda_j \boldsymbol{z}_j\boldsymbol{z}_j^\top. \tag{2.3}$$

Since we imposed the simplifying assumption that all the vectors $(\boldsymbol{z}_i)_{i=1}^p$ are independent of each other, we see that $\boldsymbol{A}$ is "almost independent" from $\boldsymbol{z}_i$. Indeed, only one term in



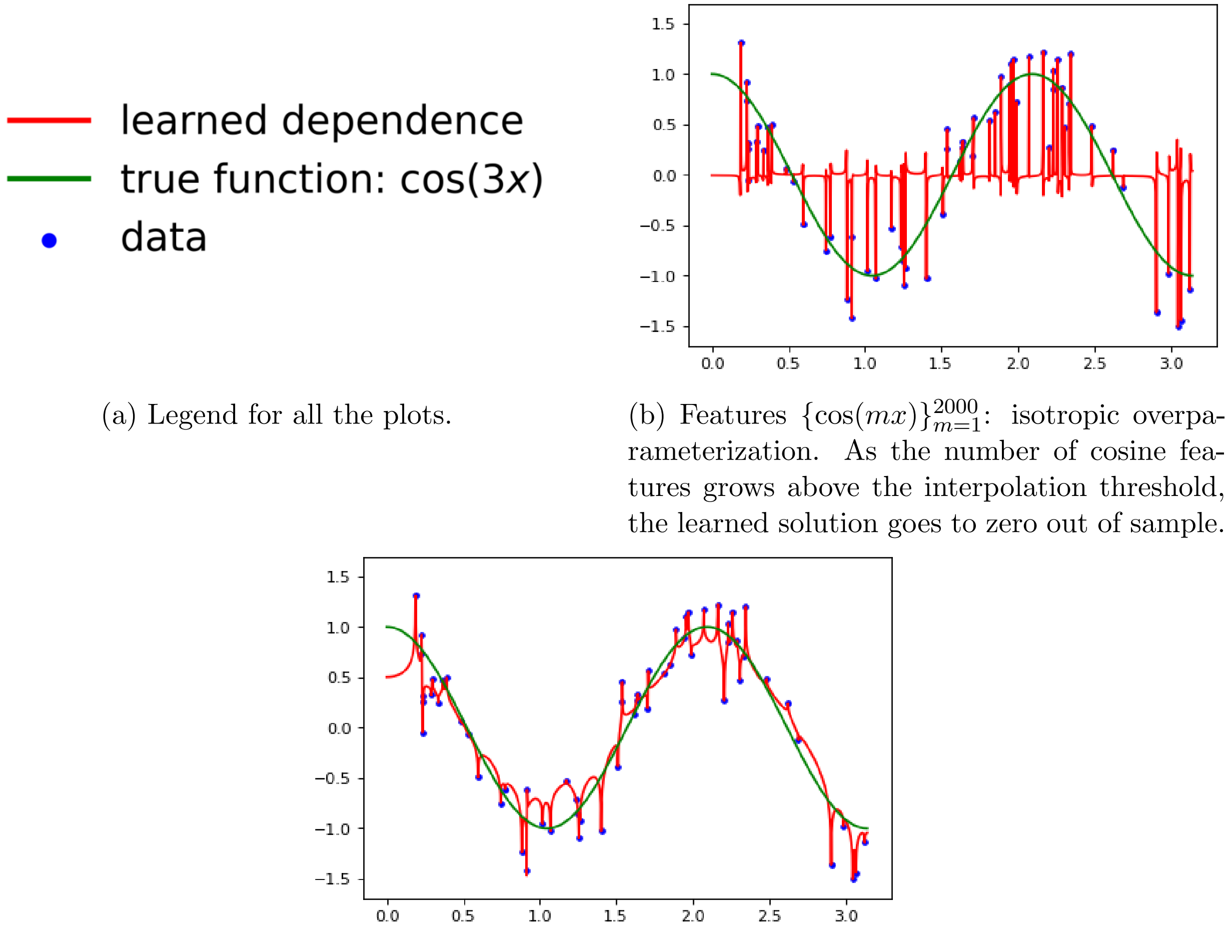
(a) Legend for all the plots.

(b) Features $\{\cos(mx)\}_{m=1}^{2}$: underfitting. A linear combination of features cannot approximate the true dependence.

(c) Features $\{\cos(mx)\}_{m=1}^{3}$: the best fit. This is the minimum number of features that span the true dependence.

(d) Features $\{\cos(mx)\}_{m=1}^{50}$: overfitting. As the number of features approaches the number of data points, the effect of the noise becomes stronger.

Figure 2.1: Learning $\cos(3x)$ using linear regression with different featurizations. The number of features is less than the number of data points, and the OLS estimator is used. The data points $(x_i, y_i)_{i=1}^{60}$ were generated i.i.d. such that x_i have uniform distribution on $[0, \pi]$ and y_i have normal distribution with mean $\cos(3x_i)$ and standard deviation 0.4.



(a) Legend for all the plots.

(b) Features $\{\cos(mx)\}_{m=1}^{2000}$: isotropic overparameterization. As the number of cosine features grows above the interpolation threshold, the learned solution goes to zero out of sample.

(c) Features $\{\cos(mx)/\sqrt{m}\}_{m=1}^{2000}$: benign overfitting. Adding weights to cosine features results in interpolating the noise with high frequency features and learning the signal with low frequency features.

Figure 2.2: Learning $\cos(3x)$ using linear regression with different featurizations. The number of data points is lower than the number of features, and the minimum norm interpolating solution is used. The data points $(x_i, y_i)_{i=1}^{60}$ were generated i.i.d. such that x_i have uniform distribution on $[0, \pi]$ and y_i have normal distribution with mean $\cos(3x_i)$ and standard deviation 0.4.

Equation 2.3 is not independent of $\boldsymbol{z}_i$: the one for which $j = i$. That is, $\boldsymbol{A}$ is a rank one correction to a matrix that is independent from $\boldsymbol{z}_i$, which gives us the next idea: use Sherman–Morrison formula to disentangle $\boldsymbol{A}$ and $\boldsymbol{z}_i$ for every i. This leads to the following lemma, whose proof can be found in Appendix A.2.

Lemma 3. *For any $i \in \{1, \dots, p\}$ define $\boldsymbol{A}_{-i} := \sum_{j \neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top$. If $\boldsymbol{A}_{-i}$ is invertible, then*

$$\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}^{-2} \boldsymbol{z}_i = \frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2}.$$

Now quadratic form concentration gives us that $\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i \approx \operatorname{tr}(\boldsymbol{A}_{-i}^{-2})$ and $\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i \approx \operatorname{tr}(\boldsymbol{A}_{-i}^{-1})$, and thus the next step is to understand the values of those traces. Since we haven't restricted the sequence $(\lambda_i)_{i=1}^p$ in any way so far, the eigenvalues of all the matrices $\boldsymbol{A}_{-i}$, as well as $\boldsymbol{A}$, can be approached in a completely analogous way. Because of that, let's focus on the matrix $\boldsymbol{A}$. The following lemma is a result of a straightforward application of an epsilon-net argument to matrix $\boldsymbol{A}$ using Equation (2.3). See Appendix A.2 for the proof.

Lemma 4. *Set $\lambda = 0$. Suppose all elements of matrix $\boldsymbol{Z}$ are independent and σ_x-sub-Gaussian. There is a constant c that only depends on σ_x such that with probability* $1 - 2e^{-n}$

$$\mu_n(\boldsymbol{A}) \geq \sum_{i=1}^p \lambda_i - c\left(n\lambda_1 + \sqrt{n \sum_{i=1}^p \lambda_i^2}\right) = tr(\boldsymbol{\Sigma}) - c(n\|\boldsymbol{\Sigma}\| + \sqrt{n}\|\boldsymbol{\Sigma}\|_F),$$

$$\mu_1(\boldsymbol{A}) \leq \sum_{i=1}^p \lambda_i + c\left(n\lambda_1 + \sqrt{n \sum_{i=1}^p \lambda_i^2}\right) = tr(\boldsymbol{\Sigma}) + c(n\|\boldsymbol{\Sigma}\| + \sqrt{n}\|\boldsymbol{\Sigma}\|_F).$$

One can see that Lemma 4 may give sharp bounds for the eigenvalues under the condition that $n\|\boldsymbol{\Sigma}\| \ll \operatorname{tr}(\boldsymbol{\Sigma})$. Indeed, since $n\|\boldsymbol{\Sigma}\|_F^2 \leq n\|\boldsymbol{\Sigma}\|\operatorname{tr}(\boldsymbol{\Sigma})$, that condition would also mean that $\sqrt{n}\|\boldsymbol{\Sigma}\|_F \ll \operatorname{tr}(\boldsymbol{\Sigma})$. However, imposing such an assumption on the sequence $(\lambda_i)_{i=1}^p$ doesn't seem right for the following reason: our actual goal is to estimate $\operatorname{tr}(\boldsymbol{A}^{-1})$ and $\operatorname{tr}(\boldsymbol{A}^{-2})$. Large eigenvalues of $\boldsymbol{A}$ give a small contribution to those quantities, so it is only important to estimate the small eigenvalues. If we change λ_1 to some large value, it will only yield a rank one correction to $\boldsymbol{A}$ and change its spectrum essentially by only changing the largest eigenvalue. Therefore, this operation would not change $\operatorname{tr}(\boldsymbol{A}^{-1})$ and $\operatorname{tr}(\boldsymbol{A}^{-2})$ much, but would completely destroy the assumption that $n\|\boldsymbol{\Sigma}\| \ll \operatorname{tr}(\boldsymbol{\Sigma})$. This is how the main idea of our work comes up for the first time: since we only really care about the small eigenvalues of $\boldsymbol{A}$, we can apply Lemma 4 to a low rank correction to $\boldsymbol{A}$ that has similar smallest eigenvalues. This is why for every $k \in \{0, \dots, p-1\}$ we introduce the following matrix:

$$\boldsymbol{A}_k := \sum_{j>k} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top. \tag{2.4}$$

Now Lemma 4 applied to $\boldsymbol{A}_k$ instead of $\boldsymbol{A}$ would yield sharp bounds on the eigenvalues under the condition that $n\lambda_{k+1}$ is much smaller than $\sum_{i>k}\lambda_i$. This motivates the following definition:

Definition 5 (Effective Rank)**.** *For $k \geq 0$, define the effective rank of the sequence $(\lambda_i)_{i>k}$ as*

$$r_k := \frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}}.$$

With this definition, we can formulate the following lemma about the eigenvalues of $\boldsymbol{A}_k$:

Lemma 6. *Set $\lambda = 0$. Suppose all elements of matrix $\boldsymbol{Z}$ are independent and σ_x-sub-Gaussian. There is a constant c that only depends on σ_x such that the following holds. Suppose that $r_k > c$ for some k. Then with probability $1 - 2e^{-n}$*

$$c^{-1}\sum_{i>k}\lambda_i \leq \mu_n(\boldsymbol{A}_k) \leq \mu_1(\boldsymbol{A}_k) \leq c\sum_{i>k}\lambda_i.$$

Proof. Denote the constant from Lemma 4 as c_1. By that lemma, with probability $1 - 2e^{-n}$

$$\begin{aligned}\mu_1(\boldsymbol{A}_k) \leq &\operatorname{tr}(\boldsymbol{\Sigma}_{k:\infty}) + c_1(n\|\boldsymbol{\Sigma}_{k:\infty}\| + \sqrt{n}\|\boldsymbol{\Sigma}_{k:\infty}\|_F),\\ \mu_n(\boldsymbol{A}_k) \geq &\operatorname{tr}(\boldsymbol{\Sigma}_{k:\infty}) - c_1(n\|\boldsymbol{\Sigma}_{k:\infty}\| + \sqrt{n}\|\boldsymbol{\Sigma}_{k:\infty}\|_F).\end{aligned}$$

Note that for any scalar w AM-GM inequality yields

$$\sqrt{n}\|\boldsymbol{\Sigma}_{k:\infty}\|_F =\leq \sqrt{n\|\boldsymbol{\Sigma}_{k:\infty}\|\operatorname{tr}(\boldsymbol{\Sigma}_{k:\infty})} \leq wn\|\boldsymbol{\Sigma}_{k:\infty}\| + w^{-1}\operatorname{tr}(\boldsymbol{\Sigma}_{k:\infty}).$$

Taking $w = 2c_1$ and plugging in this inequality into the bounds above gives the following:

$$\begin{aligned}\mu_1(\boldsymbol{A}_k) \leq &1.5\operatorname{tr}(\boldsymbol{\Sigma}_{k:\infty}) + (c_1 + 2c_1^2)n\|\boldsymbol{\Sigma}_{k:\infty}\| = (1.5 + (c_1 + 2c_1^2)n/r_k)\operatorname{tr}(\boldsymbol{\Sigma}_{k:\infty}),\\ \mu_n(\boldsymbol{A}_k) \geq &0.5\operatorname{tr}(\boldsymbol{\Sigma}_{k:\infty}) - (c_1 + 2c_1^2)n\|\boldsymbol{\Sigma}_{k:\infty}\| = (0.5 - (c_1 + 2c_1^2)n/r_k)\operatorname{tr}(\boldsymbol{\Sigma}_{k:\infty}).\end{aligned}$$

If c is large enough, then the condition $r_k > cn$ implies $0.5 - (c_1 + 2c_1^2)n/r_k \geq c^{-1}$ and $1.5 + (c_1 + 2c_1^2)n/r_k \leq c$, which finishes the proof. □

To return from the matrix $\boldsymbol{A}_k$ to the matrices $\boldsymbol{A}$ and $\boldsymbol{A}_{-i}$ we can use the following lemma, whose proof can be found in Appendix A.2.

Lemma 7. *1. for all $i \geq 1$,*

$$\mu_{k+1}(\boldsymbol{A}_{-i}) \leq \mu_{k+1}(\boldsymbol{A}) \leq \mu_1(\boldsymbol{A}_k),$$

2. for all $1 \leq i \leq k$,

$$\mu_n(\boldsymbol{A}) \geq \mu_n(\boldsymbol{A}_{-i}) \geq \mu_n\left(\boldsymbol{A}_k\right),$$

Let's denote $\Lambda_k := \sum_{i>k} \lambda_i$. We now know that if $r_k > cn$ then the small eigenvalues of $\boldsymbol{A}$ (and $\boldsymbol{A}_{-i}$ for[1] $i \le k$) are around Λ_k^{-1}.

Plugging Lemma 3 into the formula for the variance term gives

$$V = \sum_{i=1}^p \lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}^{-2} \boldsymbol{z}_i = \sum_{i=1}^p \frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2}.$$

On the right-hand side of this equation in the denominator either 1 or $\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i$ could be the dominating quantity. Depending on that, one of the following inequalities will be sharp up to a constant factor:

$$\frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2} \le \lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i, \quad \frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2} \le \frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2}.$$

Now our task is to understand for which i we want to apply the first inequality, and for which — the second.

Assume that $r_k > cn$ for some k which is small compared to n. Plugging in quadratic form concentration together with our results for eigenvalues gives $\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i \approx n\lambda_i \Lambda_k^{-1}$. The latter quantity decreases with i, and when $i = k+1$ it becomes $n\lambda_{k+1}\Lambda_k^{-1} = n/r_k < c^{-1}$, which is small. Therefore, 1 will dominate $n\lambda_i\Lambda_k^{-1}$. Thus, we want to apply the second inequality for all $i > k$ and we only want to apply the first inequality for the first ℓ terms, where $\ell \le k$. Doing that together with plugging in $\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i \approx n\Lambda_k^{-1}$ and $\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i \approx n\Lambda_k^{-2}$ yields the following upper bound on V:

Lemma 8. *Suppose all elements of matrix $\boldsymbol{Z}$ are independent and σ_x-sub-Gaussian. There are constants $b, c \ge 1$ that only depend on σ_x such that if $0 \le k \le n/c$, $r_k \ge bn$, and $l \le k$ then with probability at least $1 - 8e^{-n/c}$,*

$$V \le c\left(\frac{l}{n} + \frac{n\sum_{i>l}\lambda_i^2}{(\lambda_{k+1} r_k)^2}\right).$$

The formal proof of this lemma requires a few more technical steps, and can be found in Appendix A.2.

Interestingly, if we are looking for a lower bound, for the first ℓ terms we can obtain it without assuming that $r_k > cn$. The idea is that the ratio $\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i / (\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2$ is lower bounded by $1/\|\boldsymbol{z}_i\|^2 \approx 1/n$ due to Cauchy-Schwartz inequality. Thus, we can write

$$\frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2} = \left(\frac{\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i}{1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i}\right)^2 \frac{\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2} \gtrsim n^{-1}\left(\frac{1}{1 + 1/(\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)}\right)^2.$$

To lower-bound this quantity, we only need to bound $\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i$ from below, which means bounding the eigenvalues of $\boldsymbol{A}_{-i}$ from above. Application of Lemma 4 to $\boldsymbol{A}_k$ instead of

[1] Even though the relation between $\boldsymbol{A}_{-i}$ and $\boldsymbol{A}_k$ is only straightforward for $i \le k$, it turns out to be enough for the rigorous argument.

$\boldsymbol{A}$ does the job, and gives a non-vacuous upper bound on eigenvalues even without high effective rank condition (even though it may not be sharp).

This idea, combined with a few other technical steps gives the following lower bound, whose proof can be found in Appendix A.2.

Lemma 9. *Suppose all elements of matrix $\boldsymbol{Z}$ are independent and σ_x-sub-Gaussian. There is a constant c that only depends on σ_x such that for any $0 \le k \le n/c$ and any $b > 1$ with probability at least $1 - 10e^{-n/c}$,*

1. *If $r_k < bn$, then $V \ge \frac{k+1}{cb^2 n}$.*
2. *If $r_k \ge bn$, then*
$$V \ge \frac{1}{cb^2} \min_{l \le k} \left(\frac{l}{n} + \frac{b^2 n \sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k)^2} \right).$$

In particular, if all choices of $k \le n/c$ give $r_k < bn$, then $r_{n/c} < bn$ implies that with probability at least $1 - 12e^{-n/c}$, $V \ge (cb)^{-2}$—at least a constant.

Note that if $r_k > cn$ for some $k < n/c$ then our upper and lower bounds coincide up to a constant factor. It may be surprising, since the choice of k is somewhat arbitrary: there may be several choices of k that satisfy those conditions, and each of those choices potentially gives a new value for the bound. However, the freedom to choose k is somewhat illusory: Lemmas 7 and 6 show that, for any qualifying value of k, the smallest eigenvalue of $\boldsymbol{A}$ is within a constant factor of $\lambda_{k+1} r_k$. Thus, any two choices of k satisfying $k \le n/c$ and $r_k \ge bn$ must have values of $\lambda_{k+1} r_k$ within constant factors.

This observation helps choose the "right" ℓ and k. Looking at the bound, we see that decreasing ℓ by one subtracts $1/n$ but adds $n\lambda_\ell^2/(\lambda_{k+1} r_k)$. Because of that, for the optimal ℓ, it should hold
$$\frac{n\lambda_\ell^2}{(\lambda_{k+1} r_k)^2} \gtrsim \frac{1}{n} \gtrsim \frac{n\lambda_{\ell+1}^2}{(\lambda_{k+1} r_k)^2}.$$

That is, ℓ is the place where the ratio $n\lambda_\ell/(\lambda_{k+1} r_k)$ switches from being more than a constant to less than a constant. If we try $\ell = k+1$, then the ratio becomes n/r_k — smaller than a small constant, so $\ell \le k$. Since $\lambda_{k+1} r_k$ is approximately the same for all values of k for which r_k is large, however, ℓ should be smaller than all of them. On the other hand, if $\ell + 1 < k$, we can write
$$1 \gtrsim \frac{n\lambda_{\ell+1}}{\lambda_{k+1} r_k} \ge \frac{n\lambda_{\ell+1}}{\lambda_{\ell+1} r_\ell} = n/r_\ell,$$

which means that r_ℓ/n should be larger than a constant. This suggests that ℓ should be taken to be equal to the minimum value of k for which r_k/n is larger than a constant. Taking such ℓ indeed works and simplifies the bound on V, as the following lemma shows. The proof is in Appendix A.2.

Lemma 10. *For any $b \geq 1$ and $k^* := \min\{k : r_k \geq bn\}$, if $k^* < \infty$, we have*

$$\min_{l \leq k^*}\left(\frac{l}{bn} + \frac{bn\sum_{i>l}\lambda_i^2}{(\lambda_{k^*+1}r_{k^*})^2}\right) = \frac{k^*}{bn} + \frac{bn\sum_{i>k^*}\lambda_i^2}{(\lambda_{k^*+1}r_{k^*})^2} = \frac{k^*}{bn} + \frac{bn}{R_{k^*}},$$

where we introduced $R_k := \left(\sum_{i>k}\lambda_i\right)^2 / \left(\sum_{i>k}\lambda_i^2\right)$.

Finally, putting everything together yields the final form of the result:

Theorem 11. *Put $\lambda = 0$ and suppose all elements of matrix $\boldsymbol{Z}$ are independent and σ_x-sub-Gaussian. There are constants $a, b, c > 1$ that only depend on σ_x for which the following holds with probability $1 - 20e^{-n/c}$. Define*

$$k^* = \min\{k \geq 0 : r_k \geq bn\},$$

where the minimum of the empty set is defined as ∞. If $k^* \geq n/c$, then $V \geq 1/a$. Otherwise,

$$V/c \leq \frac{k^*}{n} + \frac{n}{R_{k^*}} \leq cV. \tag{2.5}$$

Proof. Take b to be the constant b from Lemma 8. Take c_u to be the constant c from Lemma 8 and c_ℓ to be the constant c from Lemma 9.

Suppose that $k^*/n < \min(c_u, c_\ell)$. Then Lemmas 8 and 9 yield that with probability $1 - 12e^{-n/c_\ell} - 8e^{-n/c_u}$

$$\frac{1}{c_\ell b^2}\min_{l \leq k}\left(\frac{l}{n} + \frac{b^2 n\sum_{i>l}\lambda_i^2}{(\lambda_{k+1}r_k)^2}\right) \leq V \leq c_u \min_{l \leq k^*}\left(\frac{l}{n} + \frac{n\sum_{i>l}\lambda_i^2}{(\lambda_{k+1}r_k)^2}\right).$$

Since $b > 1$, we can weaken the right inequality to obtain the following:

$$\frac{1}{c_\ell b}\min_{l \leq k}\left(\frac{l}{bn} + \frac{bn\sum_{i>l}\lambda_i^2}{(\lambda_{k+1}r_k)^2}\right) \leq V \leq c_u b \min_{l \leq k^*}\left(\frac{l}{bn} + \frac{bn\sum_{i>l}\lambda_i^2}{(\lambda_{k+1}r_k)^2}\right).$$

Now by Lemma 10, we get

$$\frac{1}{c_\ell b}\left(\frac{k^*}{bn} + \frac{bn}{R_{k^*}}\right) \leq V \leq c_u b\left(\frac{k^*}{bn} + \frac{bn}{R_{k^*}}\right),$$

which yields Equation (2.5) for $c = b^2\max(c_\ell, c_u)$.

Now, suppose $k^* > n/c$. Take $k = \lfloor n/c \rfloor < k^*$. Due to definition of k^*, $r_k < bn$. Thus, by Lemma 9, $V \geq (k+1)/(c_\ell b^2 n) \geq (n/c)/(c_\ell b^2 n) = 1/a$ for $a = cc_\ell b^2$. □

Inspection of the proofs shows that the "essentially low-dimensional" rate k/n comes from the first k components of the vector $\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})$,[2] and the term $\left(n\sum_{i>k^*}\lambda_i^2\right)/\left(\sum_{i>k^*}\lambda_i\right)^2$ comes

[2] Recall the notation $\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})$ introduced in Section 2.2.

from the rest of the components of $\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})$. Note that if one plugs in $\lambda_i = \lambda_j$ for all $i, j > k^*$, then it becomes $\left(n \sum_{i>k^*} \lambda_i^2\right) / \left(\sum_{i>k^*} \lambda_i\right)^2 = n/(p-k^*)$ — exactly the variance term of the "essentially high-dimensional" regime discussed earlier in Section 2.3. Thus, we can make a conclusion that the only way that an interpolating solution can damp the noise by more than a constant factor is the following: the data is such that after removing k components, it becomes "essentially high-dimensional", meaning that the effective rank of its covariance is large compared to the number of data points. The learning only happens in the first k components, and the corresponding variance in the first k components is the same as for the classical least squares. The variance in the rest of the components corresponds to the "essentially high-dimensional" case, where you cannot learn but the noise is still damped. Note, however, that we haven't completely justified that story yet, because only the variance term was bounded sharply so far. We understood when the model doesn't overfit to the noise, but we haven't yet shown what the model actually learns.

Effective rank

The notion of effective rank appears in two different places throughout the derivation Theorem 11: we need large r_k to sharply estimate the eigenvalues of $\boldsymbol{A}_k$ and the first place where r_k becomes large turns out to also give the right choice of ℓ. Because of this coincidence, the final result has such a simple form.

Due to such a prominent role of the effective ranks in the argument, it is informative to discuss which sequences of $(r_k)_{k=1}^p$ are possible. The following theorem gives an answer to this question.

Theorem 12. *Consider some positive summable sequence $\{\lambda_i\}_{i=1}^\infty$, and for any non-negative integer i denote*

$$r_i := \lambda_{i+1}^{-1} \sum_{j>i} \lambda_j.$$

Then $r_i > 1$ and $\sum_i r_i^{-1} = \infty$. Moreover, for any positive sequence $\{u_i\}$ such that $\sum_{i=0}^\infty u_i^{-1} = \infty$ and for every i $u_i > 1$, there exists a positive sequence $\{\lambda_i\}$ (unique up to constant multiplier) such that $r_i \equiv u_i$. The sequence is (a constant rescaling of)

$$\lambda_k = u_{k-1}^{-1} \prod_{i=0}^{k-2} (1 - u_i^{-1}).$$

The proof can be found in Appendix A.2.

In Lemma 10 we introduced another quantity: R_k. This quantity can be seen as another measure of the effective rank of sequence $(\lambda_i)_{i>k}$. Indeed, if we fix the value of $\sum_{i>k} \lambda_i$, then under this constraint the value of $\sum_{i>k} \lambda_i^2$ will be minimized when all λ_i are equal to each other, resulting in $R_k = p - k$, and it will be maximized when λ_{k+1} is the only non-zero element in the sequence, resulting in $R_k = 1$. Thus, R_k can be seen as a measure of how spread the energy of the covariates is between the coordinates starting from $k+1$-st. Let's add dependence on $\mathbf{\Sigma}$ to the notation of ranks, that is, write $r_k(\mathbf{\Sigma})$ and $R_k(\mathbf{\Sigma})$ instead of r_k and R_k. Then the following lemma shows that the two notions of effective rank are closely related.

Lemma 13. *$r_k(\mathbf{\Sigma}) \geq 1$, $r_k^2(\mathbf{\Sigma}) = r_k(\mathbf{\Sigma}^2)R_k(\mathbf{\Sigma})$, and $r_k(\mathbf{\Sigma}^2) \leq r_k(\mathbf{\Sigma}) \leq R_k(\mathbf{\Sigma}) \leq r_k^2(\mathbf{\Sigma})$.*

Proof. The first inequality and the equality are immediate from the definitions. Together they imply $R_k(\Sigma) \leq r_k^2(\mathbf{\Sigma})$. For the second inequality,

$$r_k(\mathbf{\Sigma}^2) = \frac{\sum_{i>k} \lambda_i^2}{\lambda_{k+1}^2} \leq \frac{\lambda_{k+1} \sum_{i>k} \lambda_i}{\lambda_{k+1}^2} = r_k(\mathbf{\Sigma}).$$

Substituting this in the equality implies $r_k(\mathbf{\Sigma}) \leq R_k(\mathbf{\Sigma})$. □

Further in this chapter we will obtain a sharp bound on the bias term and see that the signal can only be learned in the first k^* components, while in the remaining components at least a constant fraction of the signal will go into prediction error. Because of that, and our bound on the variance term, we give the following definition.

Definition 14. *We say that a sequence of covariance operators $\mathbf{\Sigma}_n$ is* benign *if*

$$\lim_{n\to\infty} \frac{r_0(\mathbf{\Sigma}_n)}{n} = \lim_{n\to\infty} \frac{k_n^*}{n} = \lim_{n\to\infty} \frac{n}{R_{k_n^*}(\mathbf{\Sigma}_n)} = 0,$$

where $k_n^* = \min\{k \geq 0 : r_k(\mathbf{\Sigma}_n) \geq bn\}$ for the constant b from Theorem 11 applied for the case $\sigma_x = 1$.

In the following theorem, which is proved in Appendix A.2, we study several examples of covariance sequences and derive for which values of parameters they are benign.

Theorem 15. *Define $\lambda_{k,n} := \mu_k(\mathbf{\Sigma}_n)$ for all k, n.*

1. *If $\lambda_{k,n} = k^{-\alpha} \ln^{-\beta}(k+1)$, then $\mathbf{\Sigma}_n$ is benign if and only if $\alpha = 1$ and $\beta > 1$.*
2. *If $\lambda_{k,n} = k^{-(1+\alpha_n)}$, then $\mathbf{\Sigma}_n$ is benign if and only if $\omega(1/n) = \alpha_n = o(1)$.*
3. *If*

$$\lambda_{k,n} = \begin{cases} k^{-\alpha} & \text{if } k \leq p_n, \\ 0 & \text{otherwise,} \end{cases}$$

then $\mathbf{\Sigma}_n$ is benign if and only if either $0 < \alpha < 1$, $p_n = \omega(n)$ and $p_n = o\left(n^{1/(1-\alpha)}\right)$ or $\alpha = 1$, $p_n = e^{\omega(\sqrt{n})}$ and $p_n = e^{o(n)}$.

4. *If*

$$\lambda_{k,n} = \begin{cases} \gamma_k + \epsilon_n & \text{if } k \leq p_n, \\ 0 & \text{otherwise,} \end{cases}$$

and $\gamma_k = \Theta(\exp(-k/\tau))$, then $\mathbf{\Sigma}_n$ is benign if and only if $p_n = \omega(n)$ and $ne^{-o(n)} = \epsilon_n p_n = o(n)$.

2.4 Main bounds for ridge regression

In this section, we complete the story of Section 2.3 by providing sharp bounds on the bias term, extending the results to the setting of ridge regression with nonzero λ, and replacing the assumption of independence of the components by a much broader sufficient condition. The notion of k^* is the main discovery of Section 2.3. Now we start with separation of the first k eigendirections right away, and show that the same split leads to a bound for the bias term that is in full alignment with the previously obtained intuitive explanation.

The central object in our analysis is the following matrix:

$$\boldsymbol{A}_k := \boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top + \lambda \boldsymbol{I}_n. \tag{2.6}$$

Note that (2.6) extends the definition of $\boldsymbol{A}_k$ that we gave in Section 2.3 to the case when λ is not zero.

The matrix $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ is the Gram matrix of the data after removing the first k components. $\boldsymbol{A}_k$ is obtained from that Gram matrix by shifting all eigenvalues by the ridge regularization parameter λ.

To take into account the effect of regularization on the notion of effective rank we introduce the following notation: for any $k \in \{0, 1, 2, \dots, p-1\}$ define

$$\rho_k := \frac{1}{n\lambda_{k+1}}\left(\lambda + \sum_{i>k}\lambda_i\right).$$

For $\lambda = 0$ we have $\rho_k = r_k/n$, where r_k is the effective rank introduced in Section 2.3. For example, k^* from that section is the first index k for which ρ_k becomes larger than a constant.

In Section 2.3, the crucial step was to show that the singular values of $\boldsymbol{A}_k$ are within a constant factor of each other for $k = k^*$: Lemma 6 showed that when the components of data vectors are independent, such control over the condition number is a consequence of high effective rank. In the remainder of this chapter, the roles of effective rank and condition number of $\boldsymbol{A}_k$ are reversed. We prove sharp bounds assuming that there is some oracle that guarantees that with high probability all eigenvalues of $\boldsymbol{A}_k$ are within a constant factor of each other. Independence of components is not needed. Moreover, such control implies that ρ_k is at least a constant, which, in turn, implies sharpness of the bounds. In other words, we provide a more general condition under which the tail of the data is "essentially high dimensional" — instead of assuming independent components and high effective rank, only oracle control of condition number of $\boldsymbol{A}_k$ is needed. In Section 2.5 we provide an extensive discussion of this assumption: we show that a version of a small-ball condition for the tails of the data is required and that a stronger version of the same condition is sufficient if the data is sub-Gaussian.

The bound that we obtain for the bias term is given informally by the following expression:

$$B \approx \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2.$$

One can see how it aligns with the intuition of "essentially low-dimensional" and "essentially high-dimensional" parts: one cannot estimate the signal in the high dimensional part, so almost all of its energy $\|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}$ goes into the error. When it comes to the low-dimensional part, the high-dimensional part acts as a ridge regularizer for it, so the bias in the first k components is the same as that of ridge regression with regularization coefficient $\lambda + \sum_{i>k}\lambda_i$ (i.e., the full regularization is equal to the explicitly imposed part λ plus "implicit regularization", which is equal to the energy of the tail of the covariance.)

Our extension of the results to the ridge regression scenario allows us to answer the following question: can it happen that the "essentially high dimensional part" has too much energy, meaning that it provides too much regularization and negative λ is needed to compensate for that? In Section 2.8, we show that this indeed can happen and that the following is sufficient for it to be true: the noise and the energy of the signal in the tail (components $k:\infty$) are small compared to the signal in the spiked part (components $0:k$), but the effective rank of the tail abruptly becomes much larger than n.

The central objects in our proof are $\boldsymbol{A}_k$ and ρ_k. In principle, any control of the spectrum of $\boldsymbol{A}_k$ leads to some upper bound on B and V (see our Theorem 20), the question is when that bound is tight. The intuitive answer is the following: the bound is tight when the condition number of $\boldsymbol{A}_k$ is a constant and k is chosen correctly, meaning that either ρ_k is a constant or k is the smallest number such that ρ_k is larger than a constant (i.e., $k = k^*$).[3] Our arguments, however, only support this intuition when the following technical assumption holds for some constant $\gamma < 1$:

NoncritReg(k, γ) Assume that $\lambda > -\gamma\sum_{i>k}\lambda_i$.

The reason why this assumption is needed is that as λ approaches $-\sum_{i>k}\lambda_i$, $\mathbb{E}\boldsymbol{A}_k$ approaches zero. It still can be possible to bound the eigenvalues of $\boldsymbol{A}_k$ with high probability in such regime, but their magnitude will be smaller, and some error terms that were dominated before become significant. We do investigate such a regime in Section 2.8, where we show that negative regularization may give better rates than any value of non-negative regularization, but we only provide an upper bound there. For all the results we discuss in this section, we make Assumption *NoncritReg*(k, γ).

The focus of our work was to obtain the tight upper bound on the excess risk under minimal assumptions. Such minimal assumption turns out to be

CondNum(k, δ, L) Assume that with probability at least $1-\delta$ the matrix $\boldsymbol{A}_k$ is positive-definite (PD) with condition number at most L.

We provide a thorough discussion of this assumption in Section 2.5, for example we derive sufficient and almost matching necessary conditions for it to hold when the distribution is sub-Gaussian. The reason why we don't just assume those sufficient conditions is that we believe that sub-Gaussianity is not essential for our results to hold, as we discuss in

[3] Note that there may be several values of k that satisfy these conditions. Applying our upper bound for any of those k will yield the same result up to a constant factor.

Section 2.6. Moreover, the matrix $\boldsymbol{A}_k$ is the central object in our argument, and making an assumption on its condition number explicitly makes presentation easier.

A careful reader will notice that we have just stated that another condition is needed for the bound to be tight: k should be chosen in the right way. This, however, can be achieved by shifting k to k^* if necessary: indeed, assumptions *NoncritReg*(k, γ) and *CondNum*(k, δ, L) imply a constant lower bound on ρ_k (see Corollary 21). That means that either ρ_k is a constant, or it is more than a constant, i.e., $k > k^*$. In the latter case one can shift from k to k^* meaning that Assumption *CondNum*(k^*, δ', L') also holds with modified constants δ', L' (see Lemma 26 for the exact statement). Now applying the upper bound (Corollary 21) with $k = k^*$ gives tight result, as given by the following.

Theorem 16. *Fix any constants $b > 0$, $\gamma \in [0, 1)$, $L > 0$. Denote*

$$k^* = \min\{\kappa : \rho_\kappa > b\}.$$

There exists a constant c which only depends on σ_x, b, γ, L such that the following holds: suppose NoncritReg$(\bar{k}, \gamma)$ and CondNum$(\bar{k}, \delta, L)$ are satisfied for some $\bar{k} < n/c$ and $\delta < 1 - ce^{-n/c}$. Take $k = \min(\bar{k}, k^*)$. Then with probability at least $1 - ce^{-n/c} - \delta$

$$B/c \leq \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2, \tag{2.7}$$

$$V/c \leq \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}. \tag{2.8}$$

Moreover $\rho_k \geq c^{-1}$, NoncritReg(k, γ) holds, and there exist L', c' that only depend on σ_x, b, γ, L s.t. CondNum$(k, \delta + c'e^{-n/c'}, L')$ holds.[4]

Proof. In this proof let's call any quantities that only depend on σ_x, γ, b and L "constants". First of all, if $\bar{k} \leq k^*$ then $k = \bar{k}$. Since we are given that *NoncritReg*$(\bar{k}, \gamma)$ and *CondNum*$(\bar{k}, \delta, L)$ are satisfied, we immediately get that *NoncritReg*(k, γ) and *CondNum*$(k, \delta + c'e^{-n/c'}, L')$ are satisfied with $L' = L$ and any $c' > 0$. However, if $\bar{k} > k^*$ then $k = k^*$ and by Lemma 26 *NoncritReg*(k, γ) and *CondNum*$(k, \delta + c'e^{-n/c'}, L')$ are still satisfied for some constants c', L'. Note that the larger the constants, the looser the assumptions, so we can take our final choice of c', L' to be the maximum over two cases.

Now that we know that *NoncritReg*(k, γ) and *CondNum*$(k, \delta + c'e^{-n/c'}, L')$ are satisfied, by Corollary 21, there is a constant c_1 such that $\rho_k > 1/c_1$ and with probability at least

[4] That is, the assumptions still hold if we substitute $\bar{k}$ by k, but with different L, δ. Further we will see that satisfaction of these assumptions implies tightness of the bounds for the chosen k.

$1 - c_1 e^{-n/c_1} - c' e^{-n/c'} - \delta$

$$\begin{aligned} B/c_1 \leq & \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2, \\ V/c_1 \leq & \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}. \end{aligned}$$

Taking $c \geq c_1 + c'$ gives the first part. □

Algebraically, under Assumption *CondNum*(k, δ, L) all eigenvalues of $\boldsymbol{A}_k^{-1}$ are within a constant factor of each other, so one can pull its operator norm from the expressions and obtain an upper bound without losing tightness. This strategy, however, doesn't produce lower bounds, so we derive them in a different way. Because of that, we impose different assumptions, namely

IndepCoord Assume that all elements of matrix $\boldsymbol{X}$ are independent (i.e., data vectors have independent coordinates).[5]

for the variance term, and

ExchCoord Assume that the sequence of coordinates of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ is exchangeable (any deterministic permutation of the coordinates of whitened data vectors doesn't change their distribution).

PriorSigns$(\bar{\boldsymbol{\theta}})$ Assume that $\boldsymbol{\theta}^*$ is sampled from a prior distribution in the following way: one starts with vector $\bar{\boldsymbol{\theta}}$ and flips signs of all its coordinates with probability 0.5 independently.

for the bias term. The reason we introduce them is purely technical: taking expectation over the prior signs kills the cross-terms in the expression for bias, after which we decompose bias and variance into sums with respect to individual coordinates of the predictor, and bound each term in each sum from below. The latter is possible because of the Assumptions *IndepCoord* and *ExchCoord*. We don't believe those assumptions to be necessary for our results to be tight, but because of this mismatch in assumptions, our lower bounds don't formally show that our upper bound is always tight. What they show is that one needs some specific knowledge about the distribution to obtain better bounds. We provide a more detailed discussion of the relations between those assumptions in Section 2.6. The lower bounds themselves are given by the following

[5] Recall that we fix the basis to be the eigenbasis of the covariance from the very beginning. Because of that, Assumption *IndepCoord* is stronger than the assumption that elements of $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ are independent in some basis, that is often made in Random Matrix Theory literature.

Theorem 17. *Fix any constants $b > a > 0$, $\gamma \in [0, 1)$, $L > 0$. Denote*

$$k^* = \min\{\kappa : \rho_\kappa > b\}.$$

There exists a constant c which only depends on σ_x, a, b, γ, L such that all the following hold:

1. *For any $k \in \{0, 1, \dots, k^*\}$ under assumptions IndepCoord, NoncritReg(k, γ), if $\rho_k > a$ then with probability at least $1 - 2\delta - ce^{-c/n}$*

$$V \geq \frac{1}{c}\left(\frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}\right).$$

2. *For any $k \in \{1, 2, \dots, k^*\}$ under assumptions NoncritReg(k, γ), CondNum(k, δ, L), PriorSigns$(\bar{\boldsymbol{\theta}})$ and ExchCoord, if $\rho_k > a$ then with probability at least $1 - 2\delta - ce^{-c/n}$*

$$\mathbb{E}_{\boldsymbol{\theta}^*} B \geq \frac{1}{c}\left(\|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}}\left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2\right),$$

   where $\mathbb{E}_{\boldsymbol{\theta}^*}$ denotes expectation over a random draw of $\boldsymbol{\theta}^*$ from the distribution described in assumption PriorSigns$(\bar{\boldsymbol{\theta}})$.[6]

Proof. Lemma 22 gives a lower bound for V, and Lemmas 23 and 24 give the lower bound for B. Those lower bounds have the desired probability, but different algebraic form. To bring them to the same form as the upper bounds one needs the right k to be chosen. We assumed that $\rho_k > a$. Moreover, since $k \leq k^*$ by definition of k^* we either have $\rho_k \leq b$ or $k = k^*$. In both of those cases Theorem 25 guarantees that these lower bounds are the same as what we need up to multiplicative constants that only depend on σ_x, γ, a, b and L. □

One can notice from this proof that having separate arguments for the lower bounds results in a different algebraic form of the same bound. This different form turns out to be convenient to draw explicit connections between our results and results from earlier works. We do that in Section 2.7.

2.5 Effective ranks and control of the spectrum of $\boldsymbol{A}_k$

The central assumption that we need to compute the excess risk is Assumption *CondNum*(k, δ, L), which provides control over condition number of $\boldsymbol{A}_k$. In this section we discuss when this assumption is known to be satisfied and what are the necessary conditions for it to happen.

[6] Note that under this distribution $\|\bar{\boldsymbol{\theta}}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}} = \|\boldsymbol{\theta}^*{}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}$ and $\|\bar{\boldsymbol{\theta}}_{0:k}\|_{\boldsymbol{\Sigma}^{-1}_{0:k}} = \|\boldsymbol{\theta}^*{}_{0:k}\|_{\boldsymbol{\Sigma}^{-1}_{0:k}}$ almost surely.

Effect of λ on the condition number

Recall that $\boldsymbol{A}_k = \boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top + \lambda \boldsymbol{I}_n$, so its spectrum is the shift by λ of the spectrum of $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$, the random matrix that is equal to the Gram matrix of the projected data. There are therefore three ways of establishing a constant upper bound on the condition number of $\boldsymbol{A}_k$:

1. Establish an upper bound $\bar{\mu}$ on $\mu_1(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)$ and take $\lambda > \bar{\mu}/c$ for some constant $c > 0$. In this case, the singular values of $\boldsymbol{A}_k$ are all equal to λ (and greater than $\bar{\mu}$) up to a constant multiplier.

2. Establish upper and lower bounds $\bar{\mu}$ and $\underline{\mu}$ on $\mu_1(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)$ and $\mu_n(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)$ respectively, such that $\bar{\mu}/\underline{\mu}$ is a constant. Then take $\lambda > -\underline{\mu}/c$ for some constant $c > 1$. In this case, the singular values of $\boldsymbol{A}_k$ are all equal to $\bar{\mu}$ (or $\underline{\mu}$) up to a constant multiplier.

3. Establish upper and lower bounds $\bar{\mu}$ and $\underline{\mu}$ on $\mu_1(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)$ and $\mu_n(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)$ respectively, and take $\lambda = -\underline{\mu} + \Diamond$, where $\Diamond \geq c(\bar{\mu} - \underline{\mu})$ for a constant $c > 0$. In this case, the singular values of $\boldsymbol{A}_k$ are all equal to $\Diamond$ up to a constant multiplier. This case can be substantially different from the previous case when the singular values of $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ are very well concentrated, i.e., the gap $\bar{\mu} - \underline{\mu}$ is of smaller order than $\underline{\mu}$ itself. In this case $\Diamond$ can be a smaller order term.

Our bounds are sharp when assumption *NoncritReg*(γ) is satisfied for some $\gamma < 1$, i.e., in the first and the second case above. The third case is quite rare because it requires very good concentration of the spectrum of $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$. Moreover, in this case λ is very close to the critical negative value under which it is impossible to even guarantee that $\boldsymbol{A}_k$ is PD as it becomes negative definite in expectation. We use this regime to investigate how negative regularization can improve excess risk by more than a constant factor in Section 2.8. However, we don't expect our bounds to always be sharp in this regime.

Therefore, we focus our attention on the first two cases. In Section 2.5 we discuss informally what conditions on the distribution are necessary to bound $\mu_1(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)$ and $\mu_n(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)$, and show how notions of high effective rank and norm concentration condition arise. In Section 2.5 we combine those bounds for sub-Gaussian data with the choice of λ to provide necessary and almost matching sufficient conditions for the condition number of $\boldsymbol{A}_k$ to be constant under sub-Gaussianity. In Section 2.5 we show that sub-Gaussianity is not actually required for the condition number of $\boldsymbol{A}_k$ to be controlled with high probability: Theorem 19 states that norm concentration condition and a modified version of high effective rank condition are sufficient even if the data only has bounded $4 + \boldsymbol{\varepsilon}$ moments.

Informal necessary conditions

There are several easy observations that help understand what is needed for the condition number of $\boldsymbol{A}_k$ to be bounded. In the following we use notation $\boldsymbol{U} \succeq \boldsymbol{V}$ to denote that the matrix $\boldsymbol{U}-\boldsymbol{V}$ is PSD, we use $\boldsymbol{M}[i,i]$ to denote the i-th diagonal element of the matrix $\boldsymbol{M}$, and we denote

1. The first observation is that $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top \succeq \lambda_{k+1}\boldsymbol{z}_{k+1}\boldsymbol{z}_{k+1}^\top$, where $\boldsymbol{z}_{k+1}$ is the first column of $\boldsymbol{Z}_{k:\infty}$ —a vector with n i.i.d. coordinates with unit variance. By the law of large numbers, $\|\boldsymbol{z}_{k+1}\|^2 \approx n$, meaning that $\|\lambda_{k+1}\boldsymbol{z}_{k+1}\boldsymbol{z}_{k+1}^\top\| \approx \lambda_{k+1}n$. Therefore, $\bar{\mu} \gtrsim \lambda_{k+1}n$.

2. The second observation is that the diagonal elements of $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ are squared norms of the tails of data vectors. That is, $(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)[i,i]$ are i.i.d. random variables.

   Once again, by the law of large numbers, $\operatorname{tr}(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \approx n\sum_{i>k}\lambda_i$, which implies that $\bar{\mu} \gtrsim \sum_{i>k}\lambda_i \gtrsim \underline{\mu}$. Combining it with the first observation shows that $\bar{\mu}$ and $\underline{\mu}$ can only be within a constant multiplier of each other when $\sum_{i>k}\lambda_i \geq c\lambda_{k+1}n$ for some constant c. This is exactly the high effective rank condition $\rho_k > c$ for $\lambda = 0$.

3. The third observation is that the diagonal elements of a PD matrix themselves provide bounds on the singular values:

   $$\mu_n(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \leq \min_{i\in[n]}(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)[i,i] \leq \max_{i\in[n]}(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)[i,i] \leq \mu_1(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top).$$

   Therefore, to control condition number of $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ by a constant L with probability $1-\delta$, it is necessary to guarantee that

   $$\max_i(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)[i,i]^2 \leq L\min_j(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)[i,i].$$

   Recall that we denote an independent draw of a covariate vector as $\boldsymbol{x}$. We see that n independent random draws of the random variable $\|\boldsymbol{x}_{k:\infty}\|^2$ should all lie within a constant factor of some value, meaning that the norm of the tail of a covariate vector should be within a constant factor of a fixed value with probability $(1-\delta)^{1/n}$.

Controlling condition number under sub-Gaussianity

Sub-Gaussianity of the data implies an upper bound on $\mu_1(\boldsymbol{A}_k)$, but doesn't help with $\mu_n(\boldsymbol{A}_k)$. To see this one can consider a well-known construction: take a sub-Gaussian distribution and construct another distribution in the following way: to sample from this new distribution take a vector from the old distribution and multiply it by $\sqrt{2}$ with probability $1/2$ and by zero otherwise. The new distribution is still sub-Gaussian with the same covariance, but the Gram matrix of n i.i.d. samples from it is degenerate with probability at least $1-2^{-n}$. Therefore, an additional assumption is needed to lower bound $\mu_n(\boldsymbol{A}_k)$. As we already mentioned in Section 2.5, we need norm concentration. Since sub-Gaussianity allows to bound the norm from above, it reduces to a version of the small-ball condition: $\|\boldsymbol{x}_{k:\infty}\|$ should be lower-bounded with high probability. The formal result is given by the following

Lemma 18 (Controlling $\mu_1(\boldsymbol{A}_k)/\mu_n(\boldsymbol{A}_k)$ under sub-Gaussianity)**.** *For any $\gamma \in [0, 1)$ and $\sigma_x > 0$ there exists $c > 0$ that only depends on σ_x and γ such that under Assumption NoncritReg(k, γ) the following holds: for any $L \geq 1$*

- *If $\rho_k \geq L^2$ and with probability at least $(1-\delta)^{1/n}$*
$$\lambda + \|\boldsymbol{x}_{k:\infty}\|^2 \geq \frac{c}{L}\left(\lambda + \mathbb{E}\|\boldsymbol{x}_{k:\infty}\|^2\right),$$
then with probability at least $1 - \delta - ce^{-n/c}$
$$\mu_n(\boldsymbol{A}_k) \geq L^{-1}\mu_1(\boldsymbol{A}_k).$$
- *Suppose that it is known that with probability at least $ce^{-n/c}$ $\mu_n(\boldsymbol{A}_k) \geq L^{-1}\mu_1(\boldsymbol{A}_k)$. Then $\rho_k \geq \frac{1}{cL}$ and with probability at least $\left(1 - ce^{-n/c}\right)^{1/n}$*
$$\lambda + \|\boldsymbol{x}_{k:\infty}\|^2 \geq \frac{1}{cL}\left(\lambda + \mathbb{E}\|\boldsymbol{x}_{k:\infty}\|^2\right).$$

The proof is given in Appendix A.4. One can see that both the necessary and the sufficient conditions are that ρ_k is lower bounded by a constant and a version of small-ball condition that says that the regularized squared norm of the data exceeds a constant fraction of its expectation with probability $(1-\delta)^{1/n}$. There is, however, a gap in those constants.

Heavy-tailed case

The following is a direct corollary of Theorem 2.1 from [21]

Theorem 19. *Suppose that the distribution of the tail satisfies the following two assumptions:*

1. ***Norm concentration:*** *For some $\delta \in (0, 1/n)$, $L > 1$ and $M > 0$*
$$\mathbb{P}(L^{-1} \leq \|\boldsymbol{x}_{k:\infty}\|/M \leq L) \geq 1 - \delta.$$
2. ***Heavy-tailed effective rank:*** *for some $h > 4$ denote $r_{h,k} > 0$ to be the maximum number such that for any $\boldsymbol{a} \in \mathcal{S}^{p-k-1}$ and $t > 0$*
$$\mathbb{P}\left(\frac{\sqrt{r_{h,k}}\left|\boldsymbol{a}^\top \boldsymbol{x}_{k:\infty}\right|}{M} > t\right) \leq t^{-h}.$$

There exists a constant c that only depends on h such that with probability at least $1 - cn^{1-h/4} - n\delta$

$$\begin{aligned}\mu_1(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \leq & M^2\left(L^2 + cL^2\left(n^{1-h/4} + \sqrt{\frac{n}{r_{h,k}L^2}} + \frac{n}{r_{h,k}L^2}\right)\right),\\ \mu_n(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \geq & M^2\left(L^{-2} - cL^2\left(n^{1-h/4} + \sqrt{\frac{n}{r_{h,k}L^2}} + \frac{n}{r_{h,k}L^2}\right)\right).\end{aligned}$$

Proof. First, note that by union bound with probability at least $1 - n\delta$ all the diagonal elements of the matrix $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ belong to the segment $[L^{-2}M^2, L^2M^2]$. Next, take the bound on B_k from the Case 1 of Theorem 2.1 from [21] with the following choice of their parameters: $k = N$, $\tau = 1$, $\lambda = p$, $\sigma = 1 + p/4$, $t = \sqrt{n}$. Use that bound for vectors $\sqrt{r_{h,k}}\boldsymbol{x}_{k:\infty}^i/M$. Note that that B_k is exactly the operator norm of the off-diagonal part of $r_{h,k}\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top/M^2$. □

The quantity $r_{h,k}$ that we introduced in Theorem 19 can be interpreted as a notion of effective rank for heavy tailed distributions. Indeed, one can write

$$\sqrt{r_{h,k}} = \frac{M}{\inf\left\{\tau : \forall \boldsymbol{a} \in \mathcal{S}^{p-k-1} \forall t > 0 \; \mathbb{P}\left(\left|\boldsymbol{a}^\top \boldsymbol{x}_{k:\infty}\right| / \tau > t\right) \le t^{-h}\right\}},$$

— the ratio of the typical norm of the random vector to the scale of the worst case deviations of its one-dimensional projection. This is completely analogous to our usual definition of the effective rank: $r_k = \lambda_{k+1}^{-1}\sum_{i>k}\lambda_i$. Indeed, in sub-Gaussian case $\sqrt{\sum_{i>k}\lambda_i}$ is the typical value of the norm of the vector $x_{k:\infty}$, and $\sqrt{\lambda_{k+1}}$ is up to constant the largest sub-Gaussian norm of its one-dimensional projection. We see that the conditions under which the eigenvalues of $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ are within a constant factor of each other with high probability remain the same even in the heavy-tailed case: the norm of $\|\boldsymbol{x}_{k:\infty}\|$ concentrates within a constant factor of a fixed quantity, and the heavy-tailed effective rank $r_{h,k}$ should be large compared to the number n of data points.

2.6 Structure of the proof and role of sub-Gaussianity

Upper bound

The core of our argument is Theorem 20 given below. There are two important things to note about it: first, it only requires sub-Gaussianity and matrix $\boldsymbol{A}_k$ being positive semidefinite (which always holds with probability 1 for non-negative λ). Second, its proof decomposes very clearly into two parts: an algebraic part, which only requires $\boldsymbol{A}_k$ being PD and holds with probability 1 conditionally on this event, and a probabilistic part, where standard concentration results are directly plugged into the algebraic bounds. Because of this decomposition, it is straightforward to track how the sub-Gaussianity is used and how it can be relaxed. We provide the sketch of the proof to show these details.

Theorem 20. *There exists a (large) constant c, which only depends on σ_x, s.t. for any $k < n/c$ with probability at least $1 - ce^{-n/c}$, if the matrix $\boldsymbol{A}_k$ is PD, then*

$$\begin{aligned}
B/c \leq & \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \left(1 + \frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2} + n\lambda_{k+1}\mu_1(\boldsymbol{A}_k^{-1})\left(1 + \max(0, -\lambda)\mu_1(\boldsymbol{A}_k^{-1})\right)\right) \\
& + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \left(\frac{1}{n^2\mu_n(\boldsymbol{A}_k^{-1})^2} + \frac{\lambda_{k+1}}{n}\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\left(1 + \max(0, -\lambda)\mu_1(\boldsymbol{A}_k^{-1})\right)\right), \\
V/c \leq & \frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{k}{n} + n\mu_1(\boldsymbol{A}_k^{-1})^2\sum_{i>k}\lambda_i^2.
\end{aligned}$$

Proof sketch. The full proof of Theorem 20 can be found in Section A.9 of the appendix. The following is a sketch of its derivation.

Recall the following notation: for any $\boldsymbol{y}$

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{X}^\top(\lambda\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}.$$

In Section 2.3 we introduced the notion of k^* for which the behaviour of the variance term in the first k^* coordinates is qualitatively different than in the rest of the coordinates. The argument from that section, however, relied crucially on independence of the components of the data. The main idea that allowed us to get rid of that assumption and to obtain the tight bound for the bias term was to separate the first k coordinates from the very beginning and to use some sort of uniform convergence argument in that low-dimensional subspace.

The crucial tool that allowed us to realise this idea turned out to be the following algebraic identity that we prove in Section A.6 of the appendix:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} = \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{y}.$$

This identity allows convenient access to the error in the first k coordinates (the spiked part).

The argument decomposes clearly into two parts: algebraic and probabilistic. The algebraic part is to decompose the excess risk (up to a constant multiplier) into four terms and show that the following inequalities hold on the event that the matrix $\boldsymbol{A}_k$ is PD:
(1) Bias error in the spiked part:

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*)_{0:k} - \boldsymbol{\theta}^*_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}} \leq & \frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})}\frac{\mu_1\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^{1/2}}{\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)}\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\| \\
& + \frac{\|\boldsymbol{\theta}^*_{0:k}\|_{\boldsymbol{\Sigma}^{-1}_{0:k}}}{\mu_n(\boldsymbol{A}_k^{-1})\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)}.
\end{aligned}$$

(2) Variance error in the spiked part:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}} \leq \frac{\mu_1(\boldsymbol{A}_k^{-1})^2\mathrm{tr}(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top)}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2}.$$

(3) Variance error in the tail:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{k:\infty} - \boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \le \mu_1(\boldsymbol{A}_k^{-1})^2 \mathrm{tr}(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top).$$

(4) Bias error in the tail:

$$\begin{aligned}&\frac{1}{3}\|\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*)_{k:\infty} - \boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\ \le&\|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\mu_1(\boldsymbol{A}_k^{-1})\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2\\ +&\lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})}{\mu_k(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})^2}\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}\|^2.\end{aligned}$$

The probabilistic part of the argument is to control the quantities that arise in the algebraic bound with high probability. Namely, we plug in

- Concentration of k-dimensional sample covariance with n samples: w.h.p.

$$\mu_k\left(\frac{1}{n}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right) \approx \mu_1\left(\frac{1}{n}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right) \approx 1.$$

- Concentration of norm of vectors with i.i.d. components: w.h.p.

$$\begin{aligned}\frac{1}{n}\mathrm{tr}(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top) \lesssim& k,\\ \frac{1}{n}\mathrm{tr}(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \lesssim& \sum_{i>k}\lambda_i^2,\\ \frac{1}{n}\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2 \lesssim& \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}.\end{aligned}$$

After plugging in the probabilistic bounds, the final result is obtained by a straightforward computation. □

Note that the only probabilistic statements that are used in this proof are concentration of sample covariance in dimension k and concentration of the sum of n i.i.d. random variables. The same concentration results hold with weaker assumptions, but with larger probability. For example, under rather weak moment assumptions only a linear in dimension number of samples is needed for the sample covariance matrix to concentrate within a constant factor of the population covariance, see [52] and references therein. It is also interesting to point out that the "uniform convergence" result that we mentioned in the beginning of the proof sketch is nothing but the convergence of the empirical covariance matrix $n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}$ to its expectation $\boldsymbol{I}_k$, which is exactly the uniform convergence result that gives the bound in the "essentially low-dimensional" regime from Section 2.3.

Despite the fact that the bounds of Theorem 20 apply under very general assumptions, we don't expect them to be tight if the condition number of $\boldsymbol{A}_k$ is not bounded by a constant. When some oracle control of the condition number of $\boldsymbol{A}_k$ is provided, the bound becomes the following.

Corollary 21. *Fix any constants $\gamma \in [0,1)$ and $L > 0$. There exists a constant c that only depends on σ_x, γ, L s.t. for any $k < n/c$ and $\delta < 1 - ce^{-n/c}$ under assumptions NoncritReg(k,γ) and CondNum(k,δ,L), it holds that $\rho_k > c^{-1}$, and with probability at least* $1-\delta-ce^{-n/c}$,

$$\begin{aligned} B/c \le& \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2, \\ V/c \le& \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda+\sum_{i>k}\lambda_i\right)^2}. \end{aligned}$$

Proof sketch. Assumptions *NoncritReg*(k,γ) and *CondNum*(k,δ,L) imply that all the eigenvalues of $\boldsymbol{A}_k$ are equal to $\lambda+\sum_{i>k}\lambda_i$ up to a multiplicative constant that depends on L,γ,σ_x. Plugging it into Theorem 20 gives the result. The full proof is given in Appendix A.9. □

The sub-Gaussianity is used in Corollary 21 to ensure that $\operatorname{tr}(\boldsymbol{A}_k)$ concentrates around $n\left(\lambda+\sum_{i>k}\lambda_i\right)$. Since the diagonal elements of $\boldsymbol{A}_k$ are i.i.d. random variables, the same concentration would also hold under weaker assumptions with lower but still high probability.

It is also worth mentioning that the story about "essentially high-dimensional" and "essentially low-dimensional" parts is not just an interpretation of the final result: the whole proof strategy is in accordance with it, as we explicitly separate the two parts and bound errors in them separately.

Lower bounds

Our lower bounds have a different form from the upper bounds. We show separately that they match if the condition on effective rank is satisfied. One benefit of this approach is that the lower bounds provide a different form of the same result, which allows for different analysis. We employ it in Section 2.7.

The lower bound for the variance term is given by the following lemma, whose proof is given in Appendix A.5:

Lemma 22 (Lower bound for the variance term)**.** *Fix any constant $\gamma \in [0,1)$. There exists a constant c that only depends on σ_x and γ s.t. for any $k < n/c$ under assumptions NoncritReg(k,γ) and IndepCoord w.p. at least* $1-ce^{-n/c}$

$$V \ge \frac{1}{cn}\sum_{i=1}\min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\}.$$

One can see that the assumptions under which the lower bound is proved are different from the assumptions required for the upper bound: we require independent components here. On the one hand, it means that there could be a gap between upper and lower bounds in some particular cases where one can control the condition number of $\boldsymbol{A}_k$ without independence of components. On the other hand, it means that even such strong additional assumption as independence of components does not allow the upper bounds to be improved, which suggests that those specific cases for which the bound is not tight are rare and require even stronger additional assumptions.

The most general lower bound for the bias term that we prove requires the following assumption

StableLowEig(k, δ, L) Assume that for any $j \in \{1, 2, \dots, p\}$ with probability[7] at least $1 - \delta$

$$\mu_n(\boldsymbol{A}_{-j}) \geq \mu_n(\mathbb{E}\boldsymbol{A}_k)/L = \left(\sum_{i>k} \lambda_i + \lambda\right)/L,$$

and that $\lambda > -\sum_{i>k} \lambda_i$.

Then the bound is given by the following lemma, whose proof is given in Appendix A.5

Lemma 23 (Lower bound for the bias term)**.** *Fix any constant $L > 0$. There exists c that only depends on σ_x and L s.t. for any $k \in \{1, 2, \dots, p\}$ under assumptions PriorSigns($\bar{\boldsymbol{\theta}}$) and StableLowEig($k, \delta, L$) w.p. at least $1 - 2\delta - ce^{-n/c}$*

$$\mathbb{E}_{\boldsymbol{\theta}^*} B \geq \frac{1}{c} \sum_i \frac{\lambda_i \bar{\boldsymbol{\theta}}_i^2}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2},$$

where $\mathbb{E}_{\boldsymbol{\theta}^*}$ denotes the expectation over the random draw of $\boldsymbol{\theta}^*$ from the prior distribution described in assumption PriorSigns($\bar{\boldsymbol{\theta}}$).

Assumptions *StableLowEig*(k, δ, L) and *CondNum*(k, δ, L) are formally incomparable, but informally if $k \geq 1$ then *StableLowEig*(k, δ, L) is weaker: indeed, the matrix $\boldsymbol{A}_{-i}$ is obtained from the matrix $\boldsymbol{A}$ by subtracting $\lambda_i \boldsymbol{z}_i^\top \boldsymbol{z}_i$, while the matrix $\boldsymbol{A}_k$ is obtained from $\boldsymbol{A}$ by subtracting $\sum_{i=1}^k \lambda_i \boldsymbol{z}_i^\top \boldsymbol{z}_i$, i.e., the sum of k "largest" of the terms $\lambda_i \boldsymbol{z}_i^\top \boldsymbol{z}_i$. Therefore, the matrix $\boldsymbol{A}_{-i}$ is "larger" than $\boldsymbol{A}_k$, and controlling its lowest singular value should be easier. The following lemma, whose proof is given in Appendix A.5, formalizes this argument under Assumption *ExchCoord*:

Lemma 24. *For any $\gamma < 1$ there exists a constant c that only depends on γ and σ_x such that if assumptions CondNum(k, δ, L), NoncritReg(k, γ) and ExchCoord are satisfied for some $L \geq 1$ and $k \in \{1, 2, \dots, p\}$, then StableLowEig($k, \delta + 2e^{-n/c}, cL$) is also satisfied.*

[7] Note that the condition on probability is separate for every j, i.e., we don't assume that events hold simultaneously for all j.

When it comes to averaging over the prior given by the assumption *PriorSigns*$(\bar{\boldsymbol{\theta}})$, it just means that it is impossible to obtain a better lower bound without some specific knowledge of how signs of components of $\boldsymbol{\theta}^*$ interact with the probability distribution of the data.

Connecting upper and lower bounds

One slight inconvenience with our approach of imposing oracle control over the spectrum of $\boldsymbol{A}_k$ via Assumption *CondNum*(k, δ, L) is the following: what if the oracle provides control for the wrong value of k? There can in principle be many values of k for which such oracle control is possible, with not all of them giving the right point where the behaviour changes from "essentially low-dimensional" to "essentially high-dimensional". As an example, consider the isotropic setting with $p \gg n$: one can exclude any number k of components such that $p - k \gg n$ and still be able to control the condition number.

First of all, in accordance with the result of [3], the following theorem shows that the "right k" is the k that is not larger than k^*.

Theorem 25 (The lower bound is the same as the upper bound)**.** *Denote*

$$\begin{aligned}
\underline{B} &:= \sum_i \frac{\lambda_i |\theta_i^*|^2}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2}, \\
\overline{B} &:= \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2, \\
\underline{V} &:= \frac{1}{n}\sum_i \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k + 1)^2}\right\}, \\
\overline{V} &:= \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}.
\end{aligned}$$

Fix constants $a > 0$ and $b > 1/n$. There exists a constant $c > 0$ that only depends on a, b, s.t. the following holds: if either $\rho_k \in (a, b)$ or $k = \min\{\kappa : \rho_\kappa > b\}$, then

$$c^{-1} \le \underline{B} \,/\, \overline{B} \le 1, \quad c^{-1} \le \underline{V} \,/\, \overline{V} \le 1.$$

Proof. The proof is a rather straightforward comparison of pairs of sums term by term. It is given in Appendix A.9. □

Secondly, if the data is sub-Gaussian, then oracle control for any $k < n$ results in tight bounds, but with worse constants. This happens because of the following lemma.

Lemma 26 (k can be taken to be k^*)**.** *Fix any constants $\gamma \in [0, 1)$, $b > 0$, $L > 0$. Denote*

$$k^* = \min\{k : \rho_k > b\}.$$

There exist constants c, L' that only depend on σ_x, γ, b, L s.t. the following holds: suppose assumptions NoncritReg(k, γ) and CondNum(k, δ, L) hold for some $k \in [k^*, n]$. Then assumptions NoncritReg(k^*, γ) and CondNum($k^*, \delta + ce^{-n/c}, L'$) hold too.

Proof sketch. Since $k \geq k^*$, $\mu_n(\boldsymbol{A}_k)$ provides a lower bound for $\mu_n(\boldsymbol{A}_{k^*})$. When it comes to $\mu_1(\boldsymbol{A}_{k^*})$, it can be bounded with high-probability because the data is sub-Gaussian. The full proof is given in Appendix A.4. □

The role of sub-Gaussianity

As can be seen from the proof of Theorem 16, the strategy to obtain a tight bound is the following: ask the oracle to control the condition number of $\boldsymbol{A}_k$, if that k is too large, shift it to k^*, and then apply the bound from Corollary 21. In Section 2.5 we showed that if the norm $\|\boldsymbol{x}_{k:\infty}\|$ concentrates, and the effective rank $r_{h,k}$ is high enough, then the control over the condition number of $\boldsymbol{A}_k$ is possible even if we have very weak moment assumptions instead of sub-Gaussianity. Moreover, as we have discussed in the proof sketches, if we didn't shift from k to k^*, we would only need the usual concentration results such as the law of large numbers or concentration of k-dimensional empirical covariance matrix with n samples, which also hold under weak moment assumptions. Therefore, sub-Gaussianity is not essential to obtain the bound in the form given in Corollary 21, one just needs to substitute the sub-Gaussian concentration results with their heavy-tailed analogues. However it may not necessarily give a tight result unless the oracle is guaranteed to choose the appropriate k (e.g., $k = k^*$). To shift from k to k^* we also need an upper bound on $\|\boldsymbol{A}_{k^*}\|$, which we derive from sub-Gaussianity. According to Section 2.5, an analogous bound is still possible under weak moment assumptions, but additional work is required: to use Theorem 19 for $k = k^*$ one would need to obtain a high-probability upper bound on $\|\boldsymbol{x}_{k^*:\infty}\|$ under moment assumptions and to relate r_k which we use in definition of k^* to $r_{h,k}$, which is introduced in Theorem 19.

2.7 Alternative forms of the bounds and effect of increasing regularization

Alternative form of the bound and its relation to classical in-sample analysis

Theorem 25 reveals an alternative form of the bounds: when ρ_k is lower- and upper-bounded by constants or when $k = k^*$, the bounds on the bias and variance respectively become equal to the following up to a constant multiplier:

$$\tilde{B} := \sum_{i=1}^{p} \lambda_i |\theta_i^*|^2 \frac{\rho_k^2 \lambda_{k+1}^2}{(\rho_k \lambda_{k+1} + \lambda_i)^2}, \tag{2.9}$$

$$\tilde{V} := \frac{1}{n} \sum_{i=1}^{p} \frac{\lambda_i^2}{(\rho_k \lambda_{k+1} + \lambda_i)^2}. \tag{2.10}$$

These expressions closely resemble the classical expressions for the in-sample bias and variance of ridge regression. Indeed, a straightforward computation gives

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}\|\boldsymbol{X}\hat{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}^*\|^2 \\
=&\frac{1}{n}\|(\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I}_n)^{-1} - \boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{\theta}^*\|^2 + \frac{v_{\boldsymbol{\varepsilon}}^2}{n}\|\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I}_n)^{-1}\|_F^2 \\
=&\underbrace{\sum_{i=1}^{p} \hat{\lambda}_i \langle v_i, \boldsymbol{\theta}^*\rangle^2 \frac{(\lambda/n)^2}{\left(\lambda/n + \hat{\lambda}_i\right)^2}}_{\text{in-sample bias}} + \underbrace{v_{\boldsymbol{\varepsilon}}^2 \frac{1}{n}\sum_{i=1}^{p} \frac{\hat{\lambda}_i^2}{\left(\lambda/n + \hat{\lambda}_i\right)^2}}_{\text{in-sample variance}},
\end{aligned}$$

where $\{\hat{\lambda}_i\}_{i=1}^p$ are eigenvalues of the empirical covariance $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ and $\{v_i\}_{i=1}^p$ are the corresponding eigenvectors. Recall that $\rho_k\lambda_{k+1} = \left(\lambda + \sum_{i>k}\lambda_i\right)/n$. One can see that Equations (2.9)–(2.10) can be obtained from the classical equations for the in-sample risk by substituting the empirical eigenvalues with population eigenvalues and increasing the regularization level λ by $\sum_{i>k}\lambda_i$ — the energy in the tail of the covariance.

Similarly, $\tilde{B}$ has an interpretation as the bias term of ridge regression with infinite data: for $\bar{\lambda} > 0$ denote $\boldsymbol{\theta}^*_{\bar{\lambda}}$ to be the solution to the following "population ridge regression" problem:

$$\boldsymbol{\theta}^*_{\bar{\lambda}} = \operatorname{argmin}_{\boldsymbol{\theta}} \left[\mathbb{E}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \bar{\lambda}\|\boldsymbol{\theta}\|^2\right] = \left(\boldsymbol{\Sigma} + \frac{\bar{\lambda}}{n}\boldsymbol{I}_p\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\theta}^*.$$

A straightforward computation gives

$$\|\boldsymbol{\theta}^*_{\bar{\lambda}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 = \sum_i \lambda_i |\theta_i^*|^2 \frac{(\bar{\lambda}/n)^2}{(\lambda_i + \bar{\lambda}/n)^2},$$

which is equal to $\tilde{B}$ when $\bar{\lambda} = n\lambda_{k+1}\rho_k = \lambda + \sum_{i>k}\lambda_i$.

Dependence on λ

The alternative form of the bounds presented in Section 2.7 provides a convenient way to investigate the dependence on λ, which is cumbersome in the initial form because increasing λ may decrease k^*. This effect, however, is negligible when Equations (2.9)–(2.10) are considered, as demonstrated by the following lemma that we prove in Appendix A.9.

Lemma 27. *Suppose $k < n/c$ for some $c > 1$ and $k^* < k$. Then*

$$\lambda_{k+1}\rho_k \le \lambda_{k^*+1}\rho_{k*} \le \lambda_{k+1}\rho_k/(1 - b^{-1}c^{-1}).$$

Because of this lemma, any $k \in [k^*, n/c]$ gives the same result (up to a constant factor) in Equations (2.9)–(2.10). One can, therefore, start with some λ and the corresponding $k = k^*$ and then consider larger values of λ without decreasing k in Equations (2.9)–(2.10). The result will give sharp (up to a constant factor) bounds, which depend on λ as follows:

$$\begin{aligned}\tilde{B} &= \sum_i \lambda_i |\theta_i^*|^2 \frac{n^{-2}\left(\lambda + \sum_{i>k}\lambda_i\right)}{\left(n^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right) + \lambda_i\right)^2},\\ \tilde{V} &= \frac{1}{n}\sum_i \frac{\lambda_i^2}{\left(n^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right) + \lambda_i\right)^2},\end{aligned}$$

which are obtained by simply plugging in the definition of ρ_k into (2.9)–(2.10).

A particularly interesting case arises when λ is large enough that it dominates $\sum_{i>k}\lambda_i$ and all eigenvalues of $\boldsymbol{A}_k$ are equal to λ up to a constant multiplier. The corresponding result is given by the following corollary.

Corollary 28. *There is a large positive constant c that only depends on σ_x such that if*

$$\lambda > cn\lambda_{\lfloor n/c \rfloor} + 2\sum_{i > \lfloor n/c \rfloor} \lambda_i,$$

then

$$\begin{aligned}B/c \le& \sum_i \lambda_i |\theta_i^*|^2 \frac{(\lambda/n)^2}{(\lambda/n + \lambda_i)^2},\\ V/c \le& \frac{1}{n}\sum_i \frac{\lambda_i^2}{(\lambda/n + \lambda_i)^2}.\end{aligned}$$

Proof sketch. The full proof is given in Appendix A.9; the following is its outline:

1. Use Lemma 18 to control the eigenvalues of $\boldsymbol{A}_{\lfloor n/c \rfloor}$.
2. Use Theorem 16 to obtain the bounds for $k = k^*$.
3. Use Theorem 25 to convert the bounds into the form given in Equations (2.9)–(2.10).
4. Use Lemma 27 to substitute k^* back with $\lfloor n/c \rfloor$.
5. Since $\lambda > 2\sum_{i>k}\lambda_i$, λ/n is equal to $\rho_k\lambda_{k+1}$ up to a multiplicative constant.

□

Note that the statement of Corollary 28 does not require the notion of k^*.

Comparison with other results

As we saw in the previous section, the alternative form given by Equations (2.9)–(2.10) has milder dependence on the choice of k^* than our main bounds (2.7)–(2.8) and allows to compare to other results for in-sample error of ridge regression. In this section we use it to compare with more recent developments: the non-asymptotic bounds in [25] and [22].

First of all, we follow [25] and introduce the following notion of effective dimension of the problem:

$$d(\bar{\lambda}) := \sum_i \frac{\lambda_i}{\bar{\lambda} + \lambda_i},$$

where $\bar{\lambda}$ is a parameter which can informally be understood as effective level of regularization. [25] provide non-asymptotic bounds for B and V in the regime when

$$n \geq cd(\lambda/n)\log(1 + d(\lambda/n)), \tag{2.11}$$

(see their Theorem 2).[8] The simplified version of their results given in Remark 17 gives the following bounds:[9]

$$\begin{aligned} B \leq& \left(1 + \frac{c(1 + d(\lambda/n))}{n}\right)\sum_i \lambda_i|\theta_i^*|^2 \frac{(\lambda/n)^2}{(\lambda/n + \lambda_i)^2}, \\ V \leq& \frac{c}{n}\sum_i \frac{\lambda_i^2}{(\lambda/n + \lambda_i)^2}, \end{aligned}$$

where c is some constant that depends on the concentration properties of the data. This is the same as the result of Corollary 28, but with different constants. However, our Corollary 28 covers a wider range of λ if n is large enough. This follows from the following lemma, which is proven in Appendix A.9:

Lemma 29. *Suppose that $n \geq c^2 + c$ for some $c > 0$ and take*

$$\lambda = cn\lambda_{\lfloor n/c \rfloor} + 2\sum_{i > \lfloor n/c \rfloor} \lambda_i.$$

Then

$$d(\lambda/n) \geq \frac{n}{2\max(2, (c+1)^2)}.$$

Indeed, $d(\lambda/n)$ is a decreasing function of λ, and due to Lemma 29 the range of λ for which Corollary 28 is applicable when $d(\lambda/n) = O(n)$, while Equation (2.11) restricts to the range $d(\lambda/n) = O(n/\log n)$.

[8] Note that in [25], the scaling of the regularization parameter is different from ours: to express their results in our terms one needs to substitute their λ by λ/n in our notation.

[9] Note that under our assumptions, approx$(x) = 0$, where approx(x) is defined in Equation (7) in [25].

After we first posted our results, the following non-asymptotic bound for the interpolating regime (i.e., $\lambda = 0$) appeared in [22]: informally

$$|V - V_S| \le \frac{c}{n^{1/7}}, \quad |B - B_S| \le \frac{c\|\boldsymbol{\theta}^*\|^2}{n},$$

where c is a constant, V_S and B_S are defined as[10]

$$V_S := \tilde{\lambda}^{-1} \frac{\sum_i \frac{\lambda_i^2}{(1+\tilde{\lambda}^{-1}\lambda_i)^2}}{\sum_i \frac{\lambda_i}{(1+\tilde{\lambda}^{-1}\lambda_i)^2}}, \tag{2.12}$$

$$B_S := (1 + V_S) \sum_i \frac{\lambda_i |\theta_i^*|^2}{(1 + \tilde{\lambda}^{-1}\lambda_i)^2}, \tag{2.13}$$

and $\tilde{\lambda}$ is the solution to the equation $n = d(\tilde{\lambda})$. See their Definition 1 and Theorem 2 for the exact statement.[11]

Note that because of the equation for $\tilde{\lambda}$

$$\sum_i \frac{\lambda_i^2}{(\tilde{\lambda} + \lambda_i)^2} + \sum_i \frac{\tilde{\lambda}\lambda_i}{(\tilde{\lambda} + \lambda_i)^2} = \sum_i \frac{\lambda_i(\lambda_i + \tilde{\lambda})}{(\tilde{\lambda} + \lambda_i)^2} = d(\tilde{\lambda}) = n.$$

This allows us to rewrite (2.12)–(2.13) as

$$V_S := \frac{1}{1 - \frac{1}{n}\sum_i \frac{\lambda_i^2}{(\tilde{\lambda}+\lambda_i)^2}} \cdot \frac{1}{n} \sum_i \frac{\lambda_i^2}{(\tilde{\lambda} + \lambda_i)^2}, \tag{2.14}$$

$$B_S := (1 + V_S) \sum_i \lambda_i |\theta_i^*|^2 \frac{\tilde{\lambda}^2}{(\tilde{\lambda} + \lambda_i)^2}. \tag{2.15}$$

Comparing these equations with (2.9)–(2.10) reveals that they are the same up to a constant multiplier whenever $\tilde{V} \le 1 - 1/c$ for some constant c and $\rho_k \lambda_{k+1}$ is up to a constant equal to $\tilde{\lambda}$. In the following, we show that this is indeed the case.

Recall that these results from [22] are for the interpolating regime, i.e., $\lambda = 0$. Let's see how $\tilde{\lambda}$ is related to $\lambda_{k+1}\rho_k$. The connection is given by the following lemma.

Lemma 30. *Suppose that $k < n/c$ and $\rho_k > c$ for some constant $c > 1$. Then*

$$\frac{\tilde{\lambda}}{\lambda_{k+1}\rho_k} \in \left(1 - \frac{1}{c}, \frac{1}{1 - \frac{1}{c}}\right).$$

[10] Here we introduce the notation $\tilde{\lambda} := (\gamma c_0)^{-1}$, where γ and c_0 are parameters used in [22].

[11] Note that there is a typo in their definition of $\mathscr{V}$: a multiplicative factor of c_0 is missing.

Proof. Denote $a = \frac{\tilde{\lambda}}{\lambda_{k+1}\rho_k}$. Then we can write

$$n = \sum_i \frac{\lambda_i}{\lambda_i + a\lambda_k\rho_k} \geq \sum_{i>k} \frac{\lambda_i}{\lambda_{k+1}(a\rho_k + 1)} = \frac{n\rho_k}{a\rho_k + 1},$$

which implies $a\rho_k + 1 \geq \rho_k$, so $a \geq 1 - 1/\rho_k > 1 - 1/c$.

For the upper bound on a we write

$$n = \sum_i \frac{\lambda_i}{\lambda_i + a\lambda_k\rho_k} \leq k + \sum_{i>k} \frac{\lambda_i}{a\lambda_{k+1}\rho_k} = k + \frac{n}{a},$$

which gives $a \leq n/(n-k) < c/(c-1)$. □

The similarity of Equations (2.14)–(2.15) with our results should not be taken for granted, and it is actually quite surprising. As we explain in Section 2.9, the regime considered in [22] is significantly different, so it is rather unclear why the results would have the same form.

2.8 Negative regularization

The aim of this section is to find a family of regimes in which the optimal level of ridge regularization is negative. Since we are comparing different values of λ in this section, the following notation will be useful: recall that for any k

$$\rho_k(0) := \frac{1}{n\lambda_{k+1}} \sum_{i>k} \lambda_i,$$

the value of ρ_k for $\lambda = 0$. Intuitively, the components of the tail of the covariance provide regularization for the first k components, and the larger ρ_k is, the more is that regularization. Thus, one could expect that if there is an abrupt jump in the sequence $\{\rho_k(0)\}_{k=0}^p$, then that additional regularization is too large and negative λ may be optimal.

As we investigate further, a jump in $\rho_k(0)$ is indeed one of the sufficient conditions for optimality of negative regularization, but not the only one: the strength of the noise and how the signal is distributed among the principal components of the data also play an important role.

We start the discussion with several informal observations. The first observation is that V is a decreasing function of λ: indeed, $V = \text{tr}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{X}^\top\boldsymbol{A}^{-2}\boldsymbol{X}\boldsymbol{\Sigma}^{1/2})$ and increasing λ increases all eigenvalues of $\boldsymbol{A}$. Thus, negative regularization cannot help with damping the noise compared to non-negative regularization, and the noise should be sufficiently small in order for negative regularization to be beneficial.

Now let's look at the role of the signal in the tail. It contributes to error in two ways: first — the components in the tail are not getting estimated themselves, second — the signal that comes from those components acts as additional noise for estimation of the first k

components. When λ is non-negative, the error of the first type dominates the error of the second type, but negative λ can amplify the noise and result in error of the second type dominating. Therefore, the signal in the tail also needs to be sufficiently small in order for negative regularization to be optimal.

The final observation is the following: since we only compute the bounds up to a constant multiplier, the bound in Theorem 16 cannot distinguish between negative and zero regularization. To see this, consider the form of the bound given in Section 2.7: up to a constant factor the bound is a weighted combination in each component with weight $\lambda + \sum_{i>k} \lambda_i$, and as λ increases there is no need to change k. Now it is easy to see that for all λ in range from $-\gamma \sum_{i>k} \lambda_i$ to zero, that weight is the same up to a constant factor. Thus, negative regularization can only decrease the excess risk by more than a constant factor in the critical regime, i.e., $\lambda = -\sum_{i>k} \lambda_i + \Diamond$ where $\Diamond$ is of smaller order than $\sum_{i>k} \lambda_i$. To consider such λ and have $\boldsymbol{A}_k$ PD we need tight concentration of eigenvalues of $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ around $\sum_{i>k} \lambda_i$. To ensure such tight control we restrict ourselves to the case of independent components, i.e., when Assumption *IndepCoord* is satisfied. In this case, the eigenvalues of $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ can be bounded according to Lemma 4, which we restate below in a slightly different form.

Lemma 31. *Under assumption IndepCoord there exists a constant c that only depends on σ_x s.t. with probability at least $1 - ce^{-n/c}$,*

$$\mu_1(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \le \sum_{i>k} \lambda_i + c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right),$$

$$\mu_n(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \ge \sum_{i>k} \lambda_i - c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

The fluctuations $n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}$ will be of smaller order than $\sum_{i>k}\lambda_i$ if $\rho_k(0)$ is larger than a constant, which is shown by the following bounds:

$$n\lambda_{k+1} = \frac{1}{\rho_k(0)}\sum_{i>k}\lambda_i, \tag{2.16}$$

$$\sqrt{n\sum_{i>k}\lambda_i^2} \le \sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} = \frac{1}{\sqrt{\rho_k(0)}}\sum_{i>k}\lambda_i. \tag{2.17}$$

Using this lemma allows us to obtain following two lemmas. See Appendix A.10 for the proofs.

Lemma 32 (Lower bound on the bias for any non-negative regularization)**.** *There exist constants b, c that only depend on σ_x such that the following holds: suppose that assumptions IndepCoord and PriorSigns($\bar{\boldsymbol{\theta}}$) hold. Take $k = \min\{\kappa : \rho_\kappa(0) > b\}$ and suppose that $k > 0$. Then with probability at least $1 - ce^{-n/c}$ for any $\lambda \ge 0$*

$$\mathbb{E}_{\boldsymbol{\theta}^*} B \ge \frac{1}{c}\|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\frac{\left(\sum_{i>k}\lambda_i\right)^2}{n^2}.$$

Lemma 33 (Upper bound on excess risk for some negative regularization)**.** *There exists a constant c that only depends on σ_x such that the following holds: suppose that assumptions PriorSigns($\bar{\boldsymbol{\theta}}$) and IndepCoord hold and that $\rho_k(0) > c$ for some $k < n/c$. Assume also that*

$$v_{\boldsymbol{\varepsilon}}^2 \leq \frac{1}{c}\|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}} \frac{\left(\sum_{i>k}\lambda_i\right)^2}{n^3\left(\sum_{i>k}\lambda_i^2\right)^2}. \tag{2.18}$$

Then there exists such $\lambda < 0$ that with probability at least $1 - ce^{-n/c}$

$$\mathbb{E}_{\boldsymbol{\theta}^*}B + v_{\boldsymbol{\varepsilon}}^2 V \leq c\left(v_{\boldsymbol{\varepsilon}}^2\frac{k}{n} + v_{\boldsymbol{\varepsilon}}\|\bar{\boldsymbol{\theta}}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}\sqrt{\frac{\sum_{i>k}\lambda_i^2}{n}} + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n} + \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\right).$$

Lemma 32 provides a lower bound on the expected (over noise and $\boldsymbol{\theta}^*$) excess risk which holds w.h.p. uniformly over all non-negative λ. Lemma 33 provides an upper bound that can be achieved by some negative λ. Combining these two lemmas gives a sufficient condition for the optimal λ to be negative, which is given by the following theorem.

Theorem 34. *There exist constants b and c that only depend on σ_x such that the following holds. Suppose that assumptions PriorSigns($\bar{\boldsymbol{\theta}}$) and IndepCoord hold. Take $k = \min\{\kappa : \rho_\kappa(0) > b\}$ and suppose that $k < n/c$. The value of λ that minimizes $\mathbb{E}_{\boldsymbol{\theta}^*}B + v_{\boldsymbol{\varepsilon}}V$ will be negative with probability at least $1 - ce^{-n/c}$ if the following conditions are satisfied:*

$$\begin{aligned}
\textit{small noise:}\qquad & v_{\boldsymbol{\varepsilon}}^2 \leq \frac{\|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}}{c}\min\left(\frac{\left(\sum_{i>k}\lambda_i\right)^2}{nk}, \frac{\left(\sum_{i>k}\lambda_i\right)^4}{n^3\sum_{i>k}\lambda_i^2}\right),\\
\textit{jump in effective rank:}\qquad & \rho_k(0) > c,\\
\textit{small signal in the tail:}\qquad & \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \leq \frac{1}{c}\|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\left(\frac{\sum_{i>k}\lambda_i}{n}\right)^2.
\end{aligned}$$

Proof. It is easy to see that by taking c large enough, the conditions of Lemmas 33 and 32 are satisfied, and the upper bound in Lemma 33 becomes lower than the lower bound in Lemma 32. □

We see that the conditions indeed align with the intuition outlined in the beginning of this section: we need small variance, small signal in the tail, and a sharp jump in effective rank. However, we do not have matching lower bounds in the critical regime when Assumption *NoncritReg*(k, γ) is not satisfied for a constant $\gamma > 1$. Thus, we don't know whether these conditions are also necessary.

2.9 Comparison to related works

Motivated by the empirical success of overparametrized models, there has recently been a flurry of work aimed at understanding theoretically whether the corresponding effects can be seen in overparametrized linear regression; see, e.g., [33, 43, 5, 6, 44, 60, 62, 46] and other references in this section.

The results that aim at characterizing the generalization performance of linear methods can be split roughly into three categories. The first category is results that give exact expressions of the excess risk in the asymptotic setting with ambient dimension and the number of data points going to infinity, while their ratio goes to a constant, and the spectral density of the covariance operator converges weakly to some limiting distribution [15, 24, 59, 47].

The second category is results that make strong assumptions on the distribution of data (e.g., that data vectors have i.i.d. components or come from a uniform distribution on a sphere) and derive bounds on excess risk of linear regression with some specific features, or kernel regression with a kernel that has some specific properties [41, 18, 38, 17, 32]. Some of these results are also asymptotic, and some are non-asymptotic.

The third category is results that prove non-asymptotic bounds depending on the arbitrary structure of the covariance of the data. This is the category to which our work belongs. The other works in this category are [28], [12], [14] and [13].

There have been many related works since our results were first posted on arXiv [53], for example, [36, 37, 19, 39, 4, 9, 42, 45, 35, 50, 29, 7, 11] etc. [22] obtained a finite sample version of the asymptotic results of the old version of their paper [24]. In Section 2.7 we provide an explicit comparison with our results. More recently, [36] obtained generalization bounds for kernel ridge regression under similar assumptions to those we consider here (see their Assumption 1). [29] used the idea of separating the first k eigendirections of the covariance to study excess risk of minimum norm interpolators with arbitrary norms and Gaussian data. [4] obtained results which belong to the intersection of the first and the second categories which we described in Section 2.9 (see their Theorem 4.1). [50] constructed an example of a misspecified setting (i.e., the noise is not independent from the data) in which our results don't hold even though the condition number of the matrix $\boldsymbol{A}_k$ is a constant (see their Example 1).

Next, we provide more detailed comparison with other works and discuss some technical aspects.

Results from the first category [15, 24, 59, 47] compute exact asymptotic expressions for the excess risk assuming that p/n goes to some constant as p, n go to infinity, and that the spectral distribution of $\boldsymbol{\Sigma}$ converges to some limiting distribution. From the point of view of our approach, such distributions are indistinguishable from isotropic: indeed, the very existence of limiting spectral measure implies that almost all eigenvalues are within a constant factor of each other. Many of those works even assume explicitly that the spectrum of $\boldsymbol{\Sigma}$ is upper- and lower-bounded by two constants [47, page 7], [59, Assumption 1], [24, Theorem 3]. Our results don't need any asymptotic set up, and apply to $p = \infty$ with some fixed summable sequence λ_i, which has no meaningful notion of limiting distribution, and no separation from zero is needed. For example, our setup covers kernel regression with a fixed kernel and increasing number of data points. On the other hand, when all λ_i are within a constant factor of each other, our lower bounds become $B \geq \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}/c$ and $V \geq 1/c$, so the constant part of the whole signal doesn't get learned and the variance term is at least a constant, i.e., the asymptotic expressions obtained in the works from this category are all just different constants and our approach cannot distinguish them. Therefore, we answer significantly different questions: while the asymptotic work distinguishes between constant error rates, we investigate when the error can be less than a constant. The final difference with our work is rather technical but quite strong: all the works in this category assume that the coordinates of the data become independent if multiplied by the inverse square root of the covariance. This assumption stems from asymptotic random matrix theory techniques, on which these papers are based. To the best of our knowledge, it is not known how to extend these techniques beyond random matrices with independent elements. Our approach, however, does not require the coordinates to be independent.

When it comes to the second category, featurized or kernel regression [41, 18, 38, 17, 32], the difference from our approach is that we do not assume any particular mechanism for data generation or how the features are constructed, but we directly make assumptions about feature vectors. Our results can in principle be applied in this setting if one computes the spectrum of the population covariance for particular features or kernels and the corresponding sub-Gaussian norms. The major difficulty that precludes such a direct comparison is that that computation is not straightforward. The works from this category operate in a more particular setting and circumvent the computation of the spectrum of $\boldsymbol{\Sigma}$. On the other hand, it is not hard to trace strong similarities with our approach on the level of the proof. First of all, all the papers in this category that we are aware of assume that the data comes from a very regular distribution: either d-dimensional isotropic data with i.i.d. coordinates [32, Assumption 1], or data from the uniform distribution on the sphere [38, 17, abstracts], [41, Section 3.2], or data from the product of two uniform distributions on spheres [18, Section 2.1]. Second, in all those papers the kernel is either spherically symmetric [18, Section 2.2], [32, Equation 4] or close to being spherically symmetric due to isotropic initialization of the neural network or isotropic choice of random features [17, Assumption 1], [38, Thorem 2], [41, Section 3.2]. After that, they consider the regime where n is large compared to d^α for some α [41, Assumption 1], [18, Theorem 1], [38, 17, 32, abstracts][12]. Finally, all those papers derive that kernel regression works effectively as ridge regression with polynomial features up to degree α [41, 17, abstracts], [18, Theorem 1], [32, Proposition 1 and Section 2.3]. The only exception is [38], who derive asymptotic expressions for excess risk when the true function is affine (i.e., a polynomial of degree 1) plus Gaussian misspecification. The connection with our results is that in such a regime (uniform distribution on the sphere, spherically symmetric kernel) polynomials are exactly the eigenfunctions of the kernel operator, which plays the role of the covariance operator, and there are $k \approx d^\alpha$ of polynomials of degree at

[12] In [38] $\alpha = 1$.

most α. Thus, their approach is similar to ours: separate the first k eigendirections (or their approximations) and show that other directions act as regularization.

The third category is where our work belongs, so a more concrete comparison to other results is possible. [28] proved that negative ridge regularization is optimal in a spiked covariance model with one spike, which is a simple particular case with $k = 1$ of our results. In Section 2.8, we showed that negative regularization is optimal under a rich set of covariance structures, and gave general sufficient conditions. [12] obtain non-asymptotic bounds for bias and variance in the ridgeless setting. They assume Gaussian data and the existence of k^*, which means that our results apply in their setting. Our bound for the bias term is tight, so it cannot be worse than theirs by more than a constant multiplier. At the same time, their bound on the bias term can be much worse than ours: note that their bound depends on $\|\boldsymbol{\theta}^*\|$, while our bound scales with $\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}$, therefore their bound can be arbitrarily close to infinity while our bound stays finite. When it comes to the variance term, the bound of [12] is larger but holds with smaller probability, as they discuss when they compare their results to those in [3]. [14] start with an arbitrary covariance matrix and construct a specific data distribution for which the approximation error $\mathbb{E}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2$ has an explicit expression. We provide bounds for the excess risk $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2$, so our results are not directly comparable to theirs. [13] consider expectation of the projector on the orthogonal complement to the span of i.i.d. data with arbitrary covariance and derive tight upper and lower bounds for it with respect to Loewner order. The bias term in our setting is exactly such a projection of $\boldsymbol{\theta}^*$, but measured in $\|\cdot\|_{\boldsymbol{\Sigma}}$. Because of this mismatch in the norm, the results of [13] do not translate into our results directly, even if we consider the expectation of the bias term.

2.10 Conclusions

Our results characterize when the phenomenon of benign overfitting occurs in high dimensional linear regression, with gaussian data and more generally. We give finite sample excess risk bounds that reveal the covariance structure that ensures that the minimum norm interpolating prediction rule has near-optimal prediction accuracy. The characterization depends on two notions of the effective rank of the data covariance operator. It shows that overparameterization, that is, the existence of many low-variance and hence unimportant directions in parameter space, is essential for benign overfitting.

We then studied the excess risk of ridge regression and showed how geometry of the data can influence both which part of the signal is learned and how the noise is damped. For a range of values of the regularization parameter we showed that learning can be seen as the composition of two parts: classical ridge regression in the first k components (the "essentially low-dimensional part") and learning the zero estimator in the rest of the components (the "essentially high-dimensional part"). We introduced a general assumption under which the data is "essentially high-dimensional", and provided geometric sufficient conditions for its satisfaction. Moreover, we investigated the regime in which the "essentially high-dimensional part" has too much energy, and derived general sufficient conditions for negative regularization to be optimal: small noise, small energy of the "essentially low-dimensional part", but an abrupt jump in the effective rank.

On the technical side, our proof decouples cleanly into an algebraic part, which holds with probability 1 for non-negative regularization,[13] and the probabilistic part, where we plug in well-known concentration results from high-dimensional probability. This makes it easy to trace how different terms in the bound correspond to the parts of the estimator, and supports the geometric interpretation given above.

We provided a thorough overview of the related papers, and explained how our results are significantly different from them despite some optical similarities. Those similarities, however, are intriguing, and hint at the task of developing a unified treatment of different regimes of overparameterized linear regression as a promising direction of future work.

[13] For the case of negative regularization we need to condition on the event that all the necessary symmetric matrices are PD.

Chapter 3

Classification

3.1 Introduction

In this chapter we turn our attention to binary classification. Our main goal is to study how benign overfitting can happen in this setting.

Binary classification problem setting

We consider a mixture of two classes with the same covariances $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ and symmetric (with respect to the origin) centers $\{-\boldsymbol{\mu}, \boldsymbol{\mu}\} \subset \mathbb{R}^p$. Both classes have the same probabilities. More precisely, the matrix whose rows are the data points is given as

$$\boldsymbol{X} := \boldsymbol{y}\boldsymbol{\mu}^\top + \boldsymbol{Z}\mathbf{\Sigma}^{1/2} \in \mathbb{R}^{n \times p},$$

where $\boldsymbol{y} \in \{-1, 1\}^n$ is the vector of class labels, whose components are i.i.d. Rademacher random variables, and $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ is a matrix with i.i.d. isotropic rows. We also denote $\boldsymbol{Q} := \boldsymbol{Z}\mathbf{\Sigma}^{1/2}$ — the matrix of covariates with centers of clusters subtracted. We always consider the overparameterized regime, that is $p > n$.

We consider linear classifiers which assign label $\operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x})$ to a point $\boldsymbol{x} \in \mathbb{R}^p$. Here $\boldsymbol{w} \in \mathbb{R}^p$ is the weight vector of the classifier.

Imagine[1] that we had some control over the deviations of the rows of $\boldsymbol{Z}$ in all directions, namely, imagine that there is some function $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ such that, for any $\boldsymbol{v} \in \mathcal{S}^{p-1}$ and $t > 0$

$$\mathbb{P}(\boldsymbol{z}^\top \boldsymbol{v} < -t) \leq \phi(t),$$

where $\boldsymbol{z}^\top$ is a random draw from the same distribution as the rows of $\boldsymbol{Z}$. Then the probability of an error for the classifier $\boldsymbol{x} \to \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x})$ could be bounded as

$$\mathbb{P}\left(\boldsymbol{w}^\top(\boldsymbol{\mu} + \mathbf{\Sigma}^{1/2}\boldsymbol{z}) < 0\right) = \mathbb{P}\left(\frac{\boldsymbol{w}^\top \mathbf{\Sigma}^{1/2}}{\|\boldsymbol{w}\|_{\mathbf{\Sigma}}}\boldsymbol{z} < -\frac{\boldsymbol{w}^\top \boldsymbol{\mu}}{\|\boldsymbol{w}\|_{\mathbf{\Sigma}}}\right) \leq \phi\left(\frac{\boldsymbol{\mu}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|_{\mathbf{\Sigma}}}\right).$$

[1] We don't actually impose any assumptions on ϕ throughout the chapter.

That is, the quantity $\boldsymbol{\mu}^\top \boldsymbol{w}/\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}$ provides control over the probability of predicting the wrong label on a new data point. Moreover, if rows of $\boldsymbol{Z}$ have Gaussian distribution, then by the same argument the probability of an error is exactly $\Phi(-\boldsymbol{\mu}^\top \boldsymbol{w}/\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}})$, where Φ is the normal CDF.

The main quantitative results of this chapter are bounds on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, where $\boldsymbol{w}_{\text{ridge}}$ is the solution to the ridge regression problem defined as follows. First, we assume that we are given a vector of labels $\hat{\boldsymbol{y}}$, which contains some label-flipping noise, that is, $\hat{\boldsymbol{y}}$ is obtained from $\boldsymbol{y}$ by flipping the sign of each of its coordinates independently with probability η. Then, for a given regularization parameter $\lambda \in \mathbb{R}$ we define

$$\boldsymbol{w}_{\text{ridge}} := \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_n)^{-1} \hat{\boldsymbol{y}}, \tag{3.1}$$

where $\boldsymbol{I}_n \in \mathbb{R}^{n\times n}$ is an identity matrix. An interesting particular case of this solution arises when $\lambda = 0$. In that case $\boldsymbol{X}\boldsymbol{w} = \hat{\boldsymbol{y}}$, that is, this solution exactly interpolates labels $\hat{\boldsymbol{y}}$. When $\lambda = 0$ we introduce separate notation for $\boldsymbol{w}_{\text{ridge}}$ — $\boldsymbol{w}_{\text{MNI}}$. Here MNI stands for the Minimum Norm Interpolating solution.

Assumptions on the distribution of the covariates

When it comes to the assumptions that we impose on the data distribution, we follow the steps of Chapter 2.

First of all, let's denote the eigenvalues of $\boldsymbol{\Sigma}$ in non-increasing order as $\{\lambda_i\}_{i=1}^p$ and fix the basis to be the eigenbasis of $\boldsymbol{\Sigma}$, that is, $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$. For the remainder of the chapter we will work in this basis. We assume that for some k, which is small compared to n, removing the first k columns of the matrix $\boldsymbol{Q}$ makes the "effective rank" of its rows large compared to n. The exact notion of large effective rank that we use is somewhat technical, and we postpone its introduction to Section 3.3. When the data is Gaussian (or, more generally, if the matrix $\boldsymbol{Z}$ has independent sub-Gaussian elements), that condition would be

$$\lambda + \sum_{i>k} \lambda_i > c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right), \tag{3.2}$$

where c is a large constant. Note that the regularization parameter λ adds to the energy of the tail of the covariance spectrum $\sum_{i>k}\lambda_i$ in the left hand side of the expression. That is, the notion of effective rank depends not only on $\boldsymbol{\Sigma}$, but also on the regularization applied. In Chapter 2 we showed that for $\lambda = 0$ the condition (3.2) is necessary for benign overfitting to happen in linear regression.

Apart from the "large effective rank" condition described above, we also need several concentration inequalities to hold. Those inequalities, however, are rather standard (such as law of large numbers for i.i.d. random variables or sample covariance concentration in low dimensions), so we opt with assuming that those inequalities hold directly, instead of deriving them from assumptions on the distribution of the data. We introduce those inequalities in Section 3.3.

When we began this work on classification, our motivation to consider such a regime was rather technical: we just believed that the quantities of interest can be accurately evaluated in this regime using existing techniques. However, our results suggest that such a structure of the data is necessary for benign overfitting to occur. We elaborate on that point in Section 3.8.

First result: recovering the geometry in the noiseless setting

Even though our goal is to study benign overfitting, the first result that we obtain is actually for the "noiseless" setting, that is, $\eta = 0$ and $\boldsymbol{y} = \hat{\boldsymbol{y}}$. In this setting, our bounds show that $\boldsymbol{w}_{\text{ridge}}$ performs effectively as $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$, where $\Lambda = \lambda + \sum_{i>k}\lambda_i$, and $\boldsymbol{A} = \lambda \boldsymbol{I}_n + \boldsymbol{Q}\boldsymbol{Q}^\top$. Let's look at those two terms separately. First of all, the vector $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ has symmetric distribution and has no dependence on $\boldsymbol{\mu}$, so it plays the role of a noise term. The vector $\boldsymbol{\mu}$ should be large enough in order for that noise term not to dominate. On the other hand, the term $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ can be seen as a "ridge regularized version" of the optimal classifier. Indeed, the direction $\boldsymbol{w}$ that minimizes $\boldsymbol{\mu}^\top \boldsymbol{w}/\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}$ is $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ — $\boldsymbol{\mu}$ is multiplied by the inverse of the covariance. In the expression $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ we multiply $\boldsymbol{\mu}$ by inverse of the regularized version of the covariance, and the energy of the tail of the covariance $\sum_{i>k}\lambda_i$ adds to the explicit ridge regularization λ. Another way to think about it is to introduce k^* as it was done in Chapter 2:

$$k^* := \min\left\{k : n\lambda_{k+1} < \lambda + \sum_{i>k}\lambda_i\right\}, \quad \Lambda_* := \lambda + \sum_{i>k^*}\lambda_i.$$

Then the direction above can be rewritten up to a constant factor as

$$(n\Lambda^{-1}\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}\boldsymbol{\mu} \approx \begin{pmatrix} n^{-1}\Lambda_* \boldsymbol{\Sigma}_{0:k^*}^{-1}\boldsymbol{\mu}_{0:k^*} \\ \boldsymbol{\mu}_{k^*:\infty} \end{pmatrix}.$$

We see that the ridge regression solution performs the optimal linear transform in the first k^* coordinates, but is proportional to $\boldsymbol{\mu}$ without any transformation in the remaining coordinates. Throughout this chapter we refer to this effect as "recovering the geometry" in the first k^* components.

Therefore, we show that there are 3 regimes for the noiseless setting: when $\boldsymbol{\mu}$ is small in magnitude, the "noise term" $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ will dominate in terms of both $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ and $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, and the quantity $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ will be negative with probability close to 50%, resulting in no meaningful bound on classification performance. As the magnitude of $\boldsymbol{\mu}$ grows, the term $(\boldsymbol{\Sigma}+n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ will start dominating in terms of the scalar product with $\boldsymbol{\mu}$, while the term $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ will still dominate in terms of the norm in $\boldsymbol{\Sigma}$. This regime already yields a non-vacuous classification guarantee, but still does not exhibit the full "recovery of the geometry". Finally, as $\boldsymbol{\mu}$ becomes even larger, $\boldsymbol{w}_{\text{ridge}}$ performs effectively as $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$.

It is also worth noting that the direction $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ approaches $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ as Λ decreases, which suggests that one should always use the smallest possible (perhaps even negative) regularization to achieve the best classification performance. This conclusion is not straightforward, however, since decreasing Λ also increases the noise term $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$. Nevertheless, we show that it is indeed the case, and one cannot gain a significant (in a certain sense) increase in performance by increasing λ beyond the point at which the data has high effective rank in the tail of the covariance when $\eta = 0$.

Second result: benign overfitting

When it comes to the case with label-flipping noise, we show that the structure of the solution vector may change significantly compared to the case without that noise, depending on the magnitude of $\boldsymbol{\mu}$. In the case of MNI, adding label flipping noise multiplies the solution vector by a certain scalar, and it also picks up an additional "noise component" in the orthogonal direction, which has no dependence on $\boldsymbol{\mu}$. As $\boldsymbol{\mu}$ becomes large, that multiplicative scalar becomes close to zero-mean, that is, it flips the direction of the noiseless solution with probability close to 50%. Moreover, the new orthogonal "noise component" becomes much larger in magnitude than the noiseless solution.

Nevertheless, even though the solution vector for the noisy case may look very different from the noiseless case, the bounds on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ and $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ remain rather similar. The bound on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ remains practically the same, while the bound on $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ only picks up one additional term, corresponding to the norm in $\boldsymbol{\Sigma}$ of that additional orthogonal "noise component" that we mentioned above.

As a result, our bounds suggest[2] that the noisy solution goes through the same regimes as the noiseless one, but picks up an additional regime when $\boldsymbol{\mu}$ is very large in magnitude. In that regime, our bound completely loses dependence on $\boldsymbol{\mu}$ and becomes just a function of the covariance. Interestingly, the conditions under which the bound becomes small (that is, the conditions under which we obtain benign overfitting in this regime) are exactly the same as the conditions for benign overfitting from the Chapter 2.

Comparison of classification and regression

Let us recap the main conclusions of Chapter 2 and highlight some connections and differences with our classification setup.

In Chapter 2 we considered the minimum-norm interpolating solution for an overparameterized linear regression problem. In the notation of this chapter, it can be formulated as follows:

$$\hat{\boldsymbol{\theta}} := \operatorname{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\boldsymbol{\theta}\| \text{ s.t. } \boldsymbol{Q}\boldsymbol{\theta} = \boldsymbol{Q}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon},$$

[2] Unfortunately, we only provide a lower bound on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ for the case with label-flipping noise without a matching upper bound, so we can only make a statement about the bound itself, not the quantity $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ here. We do believe, however, that this statement applies to $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, and we speculate why in Section 3.8.

where we introduced $\boldsymbol{\theta}^* \in \mathbb{R}^p$ — the coefficients of the ground truth linear model, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ — a noise vector, which has independent centered components with variances v_ε. The random noise $\boldsymbol{\varepsilon}$ is independent from $\boldsymbol{Q}$.

The solution for $\hat{\boldsymbol{\theta}}$ has a closed form expression:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{Q}^\top \boldsymbol{A}^{-1}(\boldsymbol{Q}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}),$$

and thus the mean squared error on a test point can be bounded as

$$\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\right\|_{\boldsymbol{\Sigma}} \leq \left\|(\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\theta}^*\right\|_{\boldsymbol{\Sigma}} + \left\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{\varepsilon}\right\|_{\boldsymbol{\Sigma}}. \tag{3.3}$$

The first term in the right-hand side of Equation (3.3) constitutes the bias of the MNI solution to regression, while the second term constitutes the variance. The main result of Chapter 2 is that under the structure that we introduced in Section 1.3, the variance term becomes small for the following reason: in the first k components, the noise vector $\boldsymbol{\varepsilon}$ gets projected from dimension n onto a small dimension k, and thus its energy gets damped by a factor k/n. The remainder of that noise vector, however, gets smeared over the components $k:\infty$. Since the tail of the covariance has high effective rank, a newly sampled data point will be almost orthogonal to $\hat{\boldsymbol{\theta}}$ in those components, so it doesn't matter that they absorbed the noise. When it comes to the bias term, we showed in Chapter 2 that in components $k:\infty$ no learning happens, that is, almost all the energy of the signal $\|\boldsymbol{\theta}^*_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}$ goes into the bias term. All the learning happens in the first k components, and the corresponding part of the bias term behaves like a bias term of classical ridge regression in dimension k with regularization $\sum_{i>k}\lambda_i$. In short, the tail of the covariance provides implicit regularization to the low-dimensional linear regression in the first k components. The signal is not learned in the tail at all, but at the same time it doesn't matter that it absorbs the noise.

On the one hand, the classification setting that we consider in this chapter is fundamentally different from the regression setting. Indeed, in regression the "signal vector" $\boldsymbol{\theta}^*$ is an element of a dual space. The data matrix $\boldsymbol{Q}$ measures $\boldsymbol{\theta}^*$ through evaluating the corresponding linear function. In our classification setting the "signal vector" $\boldsymbol{\mu}$ is baked into the design matrix $\boldsymbol{X} = \boldsymbol{Q} + \boldsymbol{y}\boldsymbol{\mu}^\top$, and the matrix $\boldsymbol{Q}$ obscures $\boldsymbol{\mu}$ instead of helping to measure it. Because of that, it is not clear how to apply the high-level conclusions of the work on regression to the classification setting. For example, a naive application would suggest that if $\boldsymbol{\mu}$ is supported on the tail of the covariance (i.e. if $\|\boldsymbol{\mu}_{0:k}\| = 0$), that should result in high classification error, because no learning happens in the tail. This is not correct, as our bounds can imply arbitrarily high classification accuracy in this setting. Since $\boldsymbol{\mu}$ is baked into $\boldsymbol{X}$, a plausible interpretation could be to say that the "useful space" in which the learning happens is the span of the first k eigendirections of the covariance together with $\boldsymbol{\mu}$ itself. Unfortunately, we did not find such a decomposition of the space useful in terms of obtaining a clear argument.

On the other hand, our argument shows very strong connections to the regression setting. In Section 3.1 we stated that in the setting without label-flipping noise, $\boldsymbol{w}_{\text{ridge}}$ behaves as

$\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$, and one can immediately see that the first term in that expression is almost exactly the variance part of the regression solution. Moreover, the vector $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ arises as an approximation to the vector $n\Lambda^{-1}(\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu}$, which is directly analogous to the bias part of the regression solution. Therefore, for the setting without the label-flipping noise, the bound on $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ has almost the same expression as the bound on $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_{\boldsymbol{\Sigma}}$. Since the quantity of interest in this chapter is $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, we see (at least on a technical level) that having high classification accuracy is strongly related to having small prediction error in regression. Furthermore, in Section 3.1 we talk about "an additional noise component in the orthogonal direction". As it turns out, this component also has a very similar structure to $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ (see Section 3.2 for the precise derivation). Because of that, in the large $\boldsymbol{\mu}$ regime with the label-flipping noise, our bound on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ becomes something like $1/\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}$. That is, if $\boldsymbol{\mu}$ is large, the conditions under which the classification accuracy is high are exactly the same as the conditions under which the variance of regression is low. This suggest that the mechanism by which the regression solution "hides" the noise it interpolates is very similar to that used by the classification solution.

Overall, even though there are very concrete connections between our classification results and the regression results, identifying either a clear unifying picture of these two settings or a fundamental difference between them remains an intriguing open question.

Structure of the chapter

We start in Section 3.2 by considering MNI and providing a geometric picture explaining the structure of the solution vector. After that, in Section 3.3, we introduce the assumptions that we impose on the distribution of the data to obtain quantitative bounds. Section 3.4 gives those quantitative results, and explains the regimes they go through depending on the magnitude of $\boldsymbol{\mu}$. We put almost all the technical steps of the proofs of those bounds in the Appendix, while Section 3.5 in the main body provides their outline and points to the Appendix for those particular steps. In Section 3.6 we study the influence of ridge regularization on the error of the classifier. Finally, in Section 3.7 we provide detailed comparisons with the previous literature, and Section 3.8 concludes the chapter.

Additional notation

For any scalars a, b we denote $\min(a, b)$ by $a \wedge b$ and $\max(a, b)$ by $a \vee b$. We denote $a \vee 0$ as a_+. For any $i \in \{1, 2, \dots, p\}$ we denote the i-th coordinate vector in $\mathbb{R}^p$ as $\boldsymbol{e}_i$.

For any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ we write $\boldsymbol{u} \geq \boldsymbol{v}$ to denote that all the components of $\boldsymbol{u} - \boldsymbol{v}$ are non-negative. We use $\operatorname{diag}(\boldsymbol{v})$ to denote the diagonal matrix in $\mathbb{R}^{d\times d}$ whose diagonal elements are coordinates of $\boldsymbol{v}$. For any positive integer d we denote $\mathbf{1}_d \in \mathbb{R}^d$ to be the vector of all ones.

For a linear space $\mathcal{A} \subseteq \mathbb{R}^p$ we denote its orthogonal complement by $\mathcal{A}^\perp$.

There is a slight collision with notation because we use $\boldsymbol{\mu}$ for the centers of the clusters, and we use μ_i to denote the i-th component of $\boldsymbol{\mu}$, but we also use $\mu_i(\boldsymbol{M})$ to denote the i-th largest eigenvalue of a symmetric matrix $\boldsymbol{M}$.

We say that a random element $\boldsymbol{V}$ of some real vector space has symmetric distribution if $\boldsymbol{V}$ has the same distribution as $-\boldsymbol{V}$.

3.2 Geometric picture for minimum norm interpolation

In this section, we present a geometric view on binary classification. We restrict the discussion to MNI, that is, the ridge solution $\boldsymbol{w}_{\text{ridge}}$ with zero regularization $\lambda = 0$. We do that because minimum norm interpolation is easier to think about geometrically, while ridge regularization is a more algebraic construct.

The minimum norm interpolating solution (MNI) can be defined as

$$\boldsymbol{w}_{\text{MNI}} := \arg\min_{\mathbb{R}^p} \|\boldsymbol{w}\| \text{ s.t. } \boldsymbol{X}\boldsymbol{w} = \hat{\boldsymbol{y}},$$

that is, the vector with minimum Euclidean norm that interpolates the given labels. There is an explicit formula for it, namely $\boldsymbol{w}_{\text{MNI}} = \boldsymbol{X}^{\dagger}\hat{\boldsymbol{y}}$, where $\boldsymbol{X}^{\dagger}$ denotes the pseudo-inverse of $\boldsymbol{X}$. Unfortunately, that formula is not convenient to use since we want to decouple $\boldsymbol{\mu}$ from $\boldsymbol{Q}$. There is, however, an alternative definition: to obtain the direction of MNI, one can simply find the vector with the smallest norm in the affine span of the columns of $\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$, where we define $\boldsymbol{D}_{\hat{\boldsymbol{y}}} := \operatorname{diag}(\hat{\boldsymbol{y}})$. That is, we define

$$\tilde{\boldsymbol{w}}_{\text{MNI}} := \boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{\alpha}, \text{ where } \boldsymbol{\alpha} = \operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^n} \|\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{\alpha}\| \text{ s.t. } \boldsymbol{\alpha}^\top \mathbf{1}_n = 1.$$

In other words, $\tilde{\boldsymbol{w}}_{\text{MNI}}$ is the projection of $\mathbf{0}_p$ onto the affine span of the columns of $\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$ (recall that affine span is defined as the set of linear combinations, whose coefficients sum to one, that is, $\{\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{\alpha} \text{ s.t. } \boldsymbol{\alpha}^\top \mathbf{1}_n = 1\}$).

The precise result is given by the following proposition.

Proposition 35. *The vectors $\boldsymbol{w}_{MNI}$ and $\tilde{\boldsymbol{w}}_{MNI}$ have the same direction, but different norms. They are related to each other as follows:*

$$\boldsymbol{w}_{MNI} = \frac{\tilde{\boldsymbol{w}}_{MNI}}{\|\tilde{\boldsymbol{w}}_{MNI}\|^2}, \quad \tilde{\boldsymbol{w}}_{MNI} = \frac{\boldsymbol{w}_{MNI}}{\|\boldsymbol{w}_{MNI}\|^2}. \tag{3.4}$$

Proof. First, we can rewrite the definition of MNI as

$$\boldsymbol{w}_{\text{MNI}} = \arg\min_{\mathbb{R}^p} \|\boldsymbol{w}\| \text{ s.t. } \boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{X}\boldsymbol{w} = \mathbf{1}_n.$$

We see that $\boldsymbol{w}_{\text{MNI}}$ has the same scalar products with all the columns of $\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$, which implies that it has the same scalar product with all elements of the affine span of those columns.

Therefore, it must be perpendicular to that affine span. Note that it also lies in their linear span, and there is only one direction in that linear span that is perpendicular to the affine span: the direction of the projection of zero onto the affine span. Thus, we already obtained that $\boldsymbol{w}_{\text{MNI}}$ and $\tilde{\boldsymbol{w}}_{\text{MNI}}$ have the same direction.

When it comes to the norm, since $\tilde{\boldsymbol{w}}_{\text{MNI}}$ belongs to the affine span, ${\boldsymbol{w}_{\text{MNI}}}^\top \tilde{\boldsymbol{w}}_{\text{MNI}} = 1$. On the other hand, these vectors are colinear. This yields Equation 3.4. □

Since both those vectors have the same direction, they result in the same classification rule. We are going to start our discussion in Section 3.2, where we consider the case without label-flipping noise. As it turns out, looking at $\tilde{\boldsymbol{w}}_{\text{MNI}}$ is more convenient in that case. After that, in Section 3.2, we add label-flipping noise. There it will be more convenient to return back to $\boldsymbol{w}_{\text{MNI}}$.

MNI without label-flipping noise

We start the discussion with the case without label-flipping noise, that is $\eta = 0$ and $\hat{\boldsymbol{y}} = \boldsymbol{y}$. Introduce the following notation for the two notions of the MNI direction for clean labels:

$$\begin{aligned}
\boldsymbol{w}^c_{\text{MNI}} &:= \arg\min_{\mathbb{R}^p} \|\boldsymbol{w}\| \text{ s.t. } \boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}, \\
\tilde{\boldsymbol{w}}^c_{\text{MNI}} &:= \boldsymbol{X}^\top \boldsymbol{D}_{\boldsymbol{y}} \boldsymbol{\alpha}, \text{ where } \boldsymbol{\alpha} = \operatorname{argmin} \|\boldsymbol{X}^\top \boldsymbol{D}_{\boldsymbol{y}} \boldsymbol{\alpha}\| \text{ s.t. } \boldsymbol{\alpha}^\top \boldsymbol{y} = 1.
\end{aligned}$$

Here the superscript c stands for "clean". Plugging in the expression for $\boldsymbol{X}$ gives

$$\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}} = (\boldsymbol{Q} + \boldsymbol{y}\boldsymbol{\mu}^\top)^\top \boldsymbol{D}_{\boldsymbol{y}} = \boldsymbol{Q}^\top \boldsymbol{D}_{\boldsymbol{y}} + \boldsymbol{\mu}\boldsymbol{1}_n^\top.$$

We see that changing $\boldsymbol{\mu}$ simply shifts all the columns of $\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$ by the same vector, which gives an easy way to derive the formulas for the solution. For convenience, for the rest of this section we will explicitly track the dependence on $\boldsymbol{\mu}$ in the notation, that is, we will write $\boldsymbol{w}^c_{\text{MNI}}(\boldsymbol{\mu})$ and $\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\boldsymbol{\mu})$ instead of just $\boldsymbol{w}^c_{\text{MNI}}$ and $\tilde{\boldsymbol{w}}^c_{\text{MNI}}$.

Let's start with the case $\boldsymbol{\mu} = \boldsymbol{0}_p$, and then see how adding $\boldsymbol{\mu}$ changes things. When $\boldsymbol{\mu} = \boldsymbol{0}_p$ the matrix $\boldsymbol{X}$ coincides with $\boldsymbol{Q}$, so

$$\begin{aligned}
\boldsymbol{w}^c_{\text{MNI}}(\boldsymbol{0}_p) &= \boldsymbol{Q}^\dagger \boldsymbol{y} = \boldsymbol{Q}^\top \boldsymbol{A}^{-1} \boldsymbol{y}, \\
\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\boldsymbol{0}_p) &= \frac{\boldsymbol{w}^c_{\text{MNI}}(\boldsymbol{0}_p)}{\|\boldsymbol{w}^c_{\text{MNI}}(\boldsymbol{0}_p)\|^2} = \frac{\boldsymbol{Q}^\top \boldsymbol{A}^{-1} \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1} \boldsymbol{y}}.
\end{aligned}$$

Note that $\|\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\boldsymbol{0}_p)\|^2 = (\boldsymbol{y}^\top \boldsymbol{A}^{-1} \boldsymbol{y})^{-1}$.

As we add $\boldsymbol{\mu}$, it shifts all the columns of $\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$, and thus also their affine span. Denote the linear span of the columns of $\boldsymbol{Q}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$ as $\mathcal{Q}$, and the linear space that is parallel to the affine span of those columns as $\mathcal{Q}_A$. Note that $\mathcal{Q}_A$ is orthogonal to $\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\boldsymbol{0}_p)$, and that $\mathcal{Q} = \mathcal{Q}_A \oplus \langle \tilde{\boldsymbol{w}}^c_{\text{MNI}}(\boldsymbol{0}_p) \rangle$ — the direct sum of $\mathcal{Q}_A$ and the line spanned by $\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\boldsymbol{0}_p)$. Thus, we can decompose $\boldsymbol{\mu}$ into 3 orthogonal components: $\boldsymbol{\mu}_\perp$ — a component perpendicular to $\mathcal{Q}$,

$\boldsymbol{\mu}_{\|\mathcal{Q}_A}$ — a component lying in $\mathcal{Q}_A$, and $\boldsymbol{\mu}_{\|\boldsymbol{w}(\mathbf{0})}$ — a component in the direction of $\tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\mathbf{0}_p)$. That is,

$$\boldsymbol{\mu} = \boldsymbol{\mu}_\perp + \boldsymbol{\mu}_{\|\mathcal{Q}_A} + \underbrace{\frac{\boldsymbol{\mu}^\top \tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\mathbf{0}_p)}{\|\tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\mathbf{0}_p)\|^2}\tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\mathbf{0}_p)}_{\boldsymbol{\mu}_{\|\boldsymbol{w}(\mathbf{0})}}. \tag{3.5}$$

Recall that $\tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\boldsymbol{\mu})$ is the projection of the origin onto the affine span of the columns of $\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$. Note that $\boldsymbol{\mu}_{\|\mathcal{Q}_A}$ does not change that affine span (it shifts the affine span by a vector parallel to it). The component $\boldsymbol{\mu}_{\|\boldsymbol{w}(\mathbf{0})}$ does not change the linear span, but it shifts the affine span orthogonally, so it just gets added to that projection. Finally, $\boldsymbol{\mu}_\perp$ shifts the linear span orthogonally, so it also just gets added to that projection. Therefore, we get

$$\tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\boldsymbol{\mu}) = \tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\mathbf{0}_p) + \boldsymbol{\mu}_\perp + \frac{\boldsymbol{\mu}^\top \tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\mathbf{0}_p)}{\|\tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\mathbf{0}_p)\|^2}\tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\mathbf{0}_p).$$

Plugging in the expressions for everything we get the formula:

$$\tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\boldsymbol{\mu}) = \frac{\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}} + \underbrace{(\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu}}_{\boldsymbol{\mu}_\perp} + \frac{\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}, \tag{3.6}$$

where we introduced $\boldsymbol{\nu} := \boldsymbol{Q}\boldsymbol{\mu}$.

Now that we have this decomposition, we can discuss the quantitative implications. Recall that we are interested in the quantity $\boldsymbol{\mu}^\top \tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\boldsymbol{\mu})/\|\tilde{\boldsymbol{w}}^c_{\mathrm{MNI}}(\boldsymbol{\mu})\|_{\boldsymbol{\Sigma}}$. Thus, we compare the scalar product with $\boldsymbol{\mu}$ and the norm in $\boldsymbol{\Sigma}$ for the terms above.

Interestingly, our bounds show that the second term, $\boldsymbol{\mu}_\perp$, always dominates the third term in terms of both scalar product with $\boldsymbol{\mu}$ and the norm in $\boldsymbol{\Sigma}$ in the regime that we consider.

In Section 3.1 we claimed that in absence of label-flipping noise the solution behaves as $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$. In the case of MNI, this happens because $\boldsymbol{\mu}_\perp$ behaves like $(n\Lambda^{-1}\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$, while $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \approx n\Lambda^{-1}$. More concretely, Lemma 103 from Appendix B.7 gives $\boldsymbol{\mu}^\top \boldsymbol{\mu}_\perp \approx \boldsymbol{\mu}^\top (n\Lambda^{-1}\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ and $\|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}} \lesssim \|(n\Lambda^{-1}\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$. Moreover, tightness of the latter bound was shown in Chapter 2, since the same quantity (up to the change of notation) arises there as the bias term. We have already seen this connection to the regression setting in Section 3.1, where we also discussed that the vector $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ is directly analogous to the variance part of the minimum interpolating solution for regression.

MNI with label-flipping noise and benign overfitting

Now let's see what happens when labels are not clean and contain label-flipping noise. Recall that we denoted the linear span of the columns of $\boldsymbol{Q}^\top$ as $\mathcal{Q}$. For any $\boldsymbol{v} \in \mathbb{R}^p$ denote the projection of $\boldsymbol{v}$ on $\mathcal{Q}$ as $\boldsymbol{v}_\|$ and the projection of $\boldsymbol{v}$ on $\mathcal{Q}^\perp$ as $\boldsymbol{v}_\perp$. Note that $\boldsymbol{Q}\boldsymbol{v}_\perp = \mathbf{0}_n$ for any $\boldsymbol{v} \in \mathbb{R}^p$.

MNI interpolates labels $\hat{\boldsymbol{y}}$, that is, $(\boldsymbol{Q}+\boldsymbol{y}\boldsymbol{\mu}^\top)\boldsymbol{w}_{\text{MNI}} = \hat{\boldsymbol{y}}$. Let's consider which labels are interpolated by $\boldsymbol{w}_{\text{MNI}\perp}$ and $\boldsymbol{w}_{\text{MNI}\|}$:

$$\begin{aligned}(\boldsymbol{Q}+\boldsymbol{y}\boldsymbol{\mu}^\top)\boldsymbol{w}_{\text{MNI}\perp} =& \boldsymbol{y}\boldsymbol{\mu}^\top\boldsymbol{w}_{\text{MNI}\perp} = \alpha\boldsymbol{y},\\ (\boldsymbol{Q}+\boldsymbol{y}\boldsymbol{\mu}^\top)\boldsymbol{w}_{\text{MNI}\|} =& \hat{\boldsymbol{y}} - \alpha\boldsymbol{y},\\ \boldsymbol{Q}\boldsymbol{w}_{\text{MNI}\|} =& \hat{\boldsymbol{y}} - \alpha\boldsymbol{y} - \boldsymbol{y}\boldsymbol{\mu}^\top\boldsymbol{w}_{\text{MNI}\|} = \hat{\boldsymbol{y}} - \beta\boldsymbol{y},\end{aligned}$$

where we introduced two scalar quantities: α and β.

Note that there is a unique vector $\boldsymbol{w} \in \mathcal{Q}$ such that $\boldsymbol{Q}\boldsymbol{w} = \hat{\boldsymbol{y}} - \beta\boldsymbol{y}$, namely $\boldsymbol{w} = \boldsymbol{Q}^\dagger(\hat{\boldsymbol{y}} - \beta\boldsymbol{y})$. Thus,

$$\boldsymbol{w}_{\text{MNI}\|} = \boldsymbol{Q}^\dagger(\hat{\boldsymbol{y}} - \beta\boldsymbol{y}) = \boldsymbol{Q}^\top\boldsymbol{A}^{-1}(\hat{\boldsymbol{y}} - \beta\boldsymbol{y}).$$

On the other hand, $\boldsymbol{w}_{\text{MNI}}$ always lies in the span of the columns of $\boldsymbol{X}^\top$, and projections of those columns onto $\mathcal{Q}^\perp$ are $\pm\boldsymbol{\mu}_\perp$. Thus, $\boldsymbol{w}_{\text{MNI}\perp}$ must be collinear with $\boldsymbol{\mu}_\perp$. Therefore, for some scalars a, b

$$\boldsymbol{w}_{\text{MNI}} = \boldsymbol{Q}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} + a\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y} + b\boldsymbol{\mu}_\perp,$$

which reduces the problem to two dimensions. The next simplifying step is to move to an orthogonal basis. Since $\boldsymbol{\mu}_\perp$ is already orthogonal to $\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}$ and $\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}$, we just need to find such ξ that $\boldsymbol{Q}^\top\boldsymbol{A}^{-1}(\hat{\boldsymbol{y}} - \xi\boldsymbol{y})$ is orthogonal to $\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}$. We write

$$\begin{aligned}\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}(\hat{\boldsymbol{y}} - \xi\boldsymbol{y}) &= 0,\\ \boldsymbol{y}^\top\boldsymbol{A}^{-1}(\hat{\boldsymbol{y}} - \xi\boldsymbol{y}) &= 0,\\ \xi &= \frac{\boldsymbol{y}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}}{\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}}.\end{aligned}$$

Note that $\mathbb{E}_{\boldsymbol{y},\hat{\boldsymbol{y}}}[\boldsymbol{y}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}] = (1-2\eta)\text{tr}(\boldsymbol{A}^{-1})$ and $\mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}] = \text{tr}(\boldsymbol{A}^{-1})$, so informally $\xi \approx 1 - 2\eta$. Now denote $\tilde{\boldsymbol{y}} := \hat{\boldsymbol{y}} - \xi\boldsymbol{y}$, and we get that

$$\boldsymbol{w}_{\text{MNI}} = \boldsymbol{Q}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}} + \Delta\boldsymbol{w},$$

where $\Delta\boldsymbol{w}$ belongs to the span of $\boldsymbol{\mu}_\perp$ and $\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}$. Since $\Delta\boldsymbol{w}$ is orthogonal to $\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ and $\boldsymbol{w}_{\text{MNI}}$ is the minimum norm solution that interpolates labels $\hat{\boldsymbol{y}}$, $\Delta\boldsymbol{w}$ is the minimum norm vector that interpolates the following labels:

$$\hat{\boldsymbol{y}} - (\boldsymbol{Q}+\boldsymbol{y}\boldsymbol{\mu}^\top)\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}} = (\xi - \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}})\boldsymbol{y}.$$

These labels are just the scaling of the clean labels $\boldsymbol{y}$. Thus, $\Delta\boldsymbol{w}$ is a scaled noiseless solution, that is,

$$\boldsymbol{w}_{\text{MNI}} = \boldsymbol{Q}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}} + (\xi - \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}})\boldsymbol{w}^c_{\text{MNI}}. \tag{3.7}$$

In this expression, $\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ acts as a "noise vector": it has no dependence on $\boldsymbol{\mu}$ and it has a symmetric distribution. The term $\xi\boldsymbol{w}^c_{\text{MNI}}$ is a scaling of the noiseless solution. Recall that $\xi \approx 1 - 2\eta$, which is close to 1 when the noise level η is small. Therefore, the term

$\xi \boldsymbol{w}^c_{\text{MNI}}$ is close to the noiseless solution $\boldsymbol{w}^c_{\text{MNI}}$. The last term $-\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\boldsymbol{w}^c_{\text{MNI}}$ is also a scaling of the noiseless solution, but the scaling factor $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ has symmetric distribution, as it is a linear function of $\tilde{\boldsymbol{y}}$. That is, with probability 0.5 this term points in the opposite direction from the noiseless solution, and also seems like a "noise vector".

Now let's consider how large these terms are in Euclidean norm. To do this, it is informative to consider the scale of $\boldsymbol{\mu}$ as a parameter and to see how changing it from zero to infinity affects those Euclidean norms.

The first term, $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$, does not depend on $\boldsymbol{\mu}$, so its Euclidean norm stays the same. When it comes to $\boldsymbol{w}^c_{\text{MNI}}$, however, when $\|\boldsymbol{\mu}\| = 0$ it is equal to $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$. It is very similar to the first term, with the only difference that we substitute the "noise" part of the labels $\tilde{\boldsymbol{y}}$ by the clean labels $\boldsymbol{y}$. If we take the noise to be a small constant, those two vectors should have Euclidean norms of the same order. As $\boldsymbol{\mu}$ grows, however, the vector $\boldsymbol{w}^c_{\text{MNI}}$ starts changing, and asymptotically its norm decreases inversely proportional to the scale of $\boldsymbol{\mu}$. The third term, $-\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\boldsymbol{w}^c_{\text{MNI}}$ starts at zero because $\boldsymbol{\nu}$ scales with $\boldsymbol{\mu}$. However, because of that same scaling, as $\boldsymbol{\mu}$ grows, the vector $-\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\boldsymbol{w}^c_{\text{MNI}}$ converges to a vector of finite length. Recall that this vector is equally probable to point in the same direction as $\boldsymbol{w}^c_{\text{MNI}}$ as in the opposite direction.

Overall, we see that adding a small constant amount of label-flipping noise makes the solution look significantly different compared to the noiseless one. First, it picks up an additional scaling factor, which may potentially flip the sign of the projection on the direction of the noiseless solution when $\boldsymbol{\mu}$ is large enough. Second, it picks up an additional noise component in the orthogonal direction, whose magnitude can be comparable to or even much larger than the magnitude of the noiseless solution.

However, despite those differences, the bound for the noisy case is surprisingly similar to the bound in the noiseless case. Recall that we need to estimate two scalar quantities: $\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$ and $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}$. When it comes to the first of them, our bounds suggest[3] that the sum of $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}$ and $\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$ dominate $|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}|\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$, so $\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$ only picks up one term compared to $\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$, and that term is $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}$. When it comes to the scalar product with $\boldsymbol{\mu}$, even more cancellations occur. First, recall that for the clean solution we obtained that

$$\boldsymbol{w}^c_{\text{MNI}} = \frac{\tilde{\boldsymbol{w}}^c_{\text{MNI}}}{\|\tilde{\boldsymbol{w}}^c_{\text{MNI}}\|^2}, \quad \tilde{\boldsymbol{w}}^c_{\text{MNI}} = \frac{\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}} + \boldsymbol{\mu}_\perp + \frac{\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}.$$

Denote $S := \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\tilde{\boldsymbol{w}}^c_{\text{MNI}}\|^2$. Plugging in the formulas results in

$$\begin{aligned} S\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}} =& S(\xi \boldsymbol{\mu}^\top \boldsymbol{w}^c_{\text{MNI}} + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}(1 - \boldsymbol{\mu}^\top \boldsymbol{w}^c_{\text{MNI}})) \\ =& \xi S \boldsymbol{\mu}^\top \boldsymbol{w}^c_{\text{MNI}} + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}, \end{aligned}$$

[3] We can only make an informal statement here, since our formal arguments work with slightly different expressions: instead of $\tilde{\boldsymbol{y}}$ we use the formulas involving $\Delta\boldsymbol{y} := \hat{\boldsymbol{y}} - \boldsymbol{y}$, because it has i.i.d. components. We also don't have a proof of tightness for the case with label-flipping noise.

which already suggests that $S\boldsymbol{\mu}^\top \boldsymbol{w}_{\mathrm{MNI}}$ is similar to $S\boldsymbol{\mu}^\top \boldsymbol{w}^c_{\mathrm{MNI}}$. To see that even more clearly, however, we can plug in the formula for $S\boldsymbol{\mu}^\top \boldsymbol{w}^c_{\mathrm{MNI}}$, that is

$$S\boldsymbol{\mu}^\top \boldsymbol{w}^c_{\mathrm{MNI}} = \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}.$$

Plugging that in together with the definition of ξ gives

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{\mathrm{MNI}} = \boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\|\boldsymbol{\mu}_\perp\|^2 + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}},$$

so the formula for the scalar product in the noisy case turns out to be almost the same as in the noiseless case, and the bound doesn't change.

A part of the reason for this cancellation can be seen from our derivation of the formula: we started by saying that we need the component $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ to interpolate the "noise part of the labels" $\tilde{\boldsymbol{y}}$. When multiplied by $\boldsymbol{Q} + \boldsymbol{y}\boldsymbol{\mu}^\top$, this vector picks up additional labels proportional to $\boldsymbol{y}$ because of the scalar product with $\boldsymbol{\mu}$. We then say that the job of the additional term $\Delta\boldsymbol{w}$ is to kill those additional labels. However, the labels that $\Delta\boldsymbol{w}$ interpolates themselves come largely from its scalar product with $\boldsymbol{\mu}$. In the end, this leads to the cancellation when we compute the scalar product of $\boldsymbol{\mu}$ and the sum of $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ and $\Delta\boldsymbol{w}$. However, this only explains why some of the terms disappear. Finding an intuitive explanation why the resulting formulas for the scalar product are so similar in the noisy and noiseless cases remains an intriguing question.

3.3 Assumptions on the data

So far we have seen how MNI behaves geometrically and which terms arise in the expressions of interest. To make a quantitative statement about classification, however, one needs to bound those terms. In their turn, those bounds require assumptions.

In this section we explain which assumptions we impose on the distribution of the rows of $\boldsymbol{Z}$ and on the sequence $\{\lambda_i\}_{i=1}^p$ in order to obtain our results.

Gram matrix of the tails

The central object in our analysis is the (regularized) Gram matrix of the covariates projected onto the tail of the covariance, which we denote as follows:

$$\boldsymbol{A}_k := \boldsymbol{Q}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top + \lambda \boldsymbol{I}_n. \tag{3.8}$$

Just as in Section 2.4 of Chapter 2 the main assumption under which our argument works is that the condition number of the matrix $\boldsymbol{A}_k$ is bounded by some constant. Therefore, we introduce the following event:

Definition 36. *For any $k \in \{0, 1, \dots, p-1\}$ and $L \geq 1$ we define by $\mathscr{A}_k(L)$ the following event:*

$$\mathscr{A}_k(L) := \left\{\Big(\lambda + \sum_{i>k}\lambda_i\Big)/L \leq \mu_n(\boldsymbol{A}_k) \leq \mu_1(\boldsymbol{A}_k) \leq L\Big(\lambda + \sum_{i>k}\lambda_i\Big)\right\}. \tag{3.9}$$

Note that $\mathbb{E}\boldsymbol{A}_k = \boldsymbol{I}_n \cdot \left(\lambda + \sum_{i>k}\lambda_i\right)$, so on $\mathscr{A}_k(L)$ the eigenvalues of $\boldsymbol{A}_k$ are within a constant factor of eigenvalues of its expectation.

Throughout the chapter we will also always impose assumptions of the following form:

$$\lambda + \sum_{i>k}\lambda_i > c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right), \tag{3.10}$$

where c is some large constant. It is closely related to saying that the event $\mathscr{A}_k$ holds with high probability, and also to the notions of effective ranks used in Chapter 2. Indeed, the following lemma follows directly from Lemma 4 from Section 2.3.

Lemma 37. *Suppose that elements of $\boldsymbol{Z}$ are σ_x-sub-Gaussian and independent. There exists a constant c that only depends on σ_x s.t. with probability at least* $1 - ce^{-n/c}$,

$$\mu_1(\boldsymbol{A}_k) = \lambda + \mu_1(\boldsymbol{Q}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top) \le \lambda + \sum_{i>k}\lambda_i + c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right),$$

$$\mu_n(\boldsymbol{A}_k) = \lambda + \mu_n(\boldsymbol{Q}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top) \ge \lambda + \sum_{i>k}\lambda_i - c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

Lemma 37 shows that if components of the data are independent and sub-Gaussian, then Equation (3.10) implies that $\mathscr{A}_k(L)$ holds with high probability for some constant L. The reason why we introduce event $\mathscr{A}_k(L)$ instead of assuming independence of the components is that we don't believe that independence to be necessary for $\mathscr{A}_k(L)$ to hold. The same logic was followed in Section 2.4 of Chapter 2. Moreover, in Section 2.5 we explained that much weaker conditions, such as sub-Gaussianity and a version of a small-ball condition, are sufficient for the event $\mathscr{A}_k(L)$ to hold with high probability. We even showed that $\mathscr{A}_k(L)$ can hold with high probability for some heavy-tailed distributions. Note, however, that the strategy in Chapter 2 was slightly different: we imposed Assumption *CondNum*(k, δ, L) which directly controls the condition number of $\boldsymbol{A}_k$. Then we showed that under additional Assumption *NoncritReg*(k, γ) and sub-Gaussianity the eigenvalues of $\boldsymbol{A}_k$ are within a constant factor of $\lambda + \sum_{i>k}\lambda_i$ (see Lemma 85 in Appendix A.4). In this chapter we put direct bounds on the eigenvalues of $\boldsymbol{A}_k$ in Definition 36, which simplifies the presentation.

Algebraic assumptions

Similarly to our results on regression, our argument for the main lower bound cleanly decomposes into an algebraic and a probabilistic part. Because of that, we don't need to formulate that bound with some probability over the draw of $\boldsymbol{Q}$, but we can just specify the exact event on which our results hold. We have already introduced the event $\mathscr{A}_k(L)$. Another event that we need is as follows.

Definition 38. *For any $k \in \{0, 1, \dots, p-1\}$ and $c_B > 0$, we define by $\mathscr{B}_k(c_B)$ the event on which all the following hold:*

1. $\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \le c_B n$ *and* $\mu_n(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \ge n/c_B$.
2. $\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \le c_B n \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2$.
3. $tr(\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top) \le c_B n \sum_{i>k} \lambda_i^2$.
4. $tr(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \le c_B nk$.
5. $\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\| \le c_B \left(\sum_{i>k} \lambda_i^2 + n\lambda_{k+1}^2\right)$.

It is easy to see that the event $\mathscr{B}_k(c_B)$ holds with high probability if the constant c_B is large enough. For the case of sub-Gaussian data this can be stated more precisely:

Lemma 39. *Suppose the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. One can take the constant c_B large enough depending only on σ_x such that for any $k < n/c_B$ the probability of the event $\mathscr{B}_k(c_B)$ is at least $1 - c_B e^{-n/c_B}$.*

Proof. We need to show that all 5 bounds from the definition of $\mathscr{B}_k(c_B)$ hold with probability at least $1 - ce^{-n/c}$, where c only depends on σ_x. The first 4 of these bounds were shown in Appendix A.9 up to the change of notation; see the display (A.18). One just needs to substitute $\boldsymbol{X}$ by $\boldsymbol{Q}$ and $\boldsymbol{\theta}^*$ by $\boldsymbol{\mu}$ in that display to get the result of the lemma. The last statement follows directly from Lemma 84 by plugging in $\boldsymbol{Z\Sigma}$ instead of $\boldsymbol{X}$. □

As with the definition of the event $\mathscr{A}_k$, we introduce the event $\mathscr{B}_k(c_B)$ instead of assuming sub-Gaussianity because we believe that sub-Gaussianity is not necessary for $\mathscr{B}_k(c_B)$ to hold with high probability. This was also discussed in Section 2.6. Indeed, the first condition in the definition of $\mathscr{B}_k(c_B)$ is just concentration of sample covariance in dimension k with n data points, which is known to hold for heavy-tailed distributions (see [52] and references therein). The inequalities 2–4 are just the law of large numbers (concentration of the sum of n i.i.d. random variables). Only the inequality number 5 (bound on the norm of the Gram matrix) is somewhat less standard. Note however, that the Gram matrix has the same spectral norm as the sample covariance matrix multiplied by n. Therefore, that inequality could be obtained as a direct corollary of a dimension-free bound on the spectral norm of a sample covariance matrix. An example of a heavy-tailed result of this type can be found in [1], see their Theorem 2.

3.4 Main results

In order to formulate the results more succinctly throughout the chapter we introduce additional notation. First of all, we denote the bound on the sub-Gaussian constant of the label-flipping noise as

$$\sigma_\eta := 1/\sqrt{\ln\frac{3+\eta^{-1}}{2}}. \tag{3.11}$$

Next, for a given k we will use the following notation:

$$\begin{aligned}
\Lambda :=& \lambda + \sum_{i>k}\lambda_i,\\
V :=& n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2}n\sum_{i>k}\lambda_i^2,\\
\Delta V :=& \frac{1}{n}\wedge\frac{n\lambda_1^2}{\Lambda^2} + \frac{n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2}{\Lambda^2},\\
B :=& n^{-2}\Lambda^2\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2,\\
\lozenge^2 :=& n\Lambda^{-2}B,\\
M :=& \frac{\Lambda}{n}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + \|\boldsymbol{\mu}_{k:\infty}\|^2,\\
N :=& n\Lambda^{-1}M.
\end{aligned}$$

Note that we don't track the dependence on k in the notation. We always introduce k before using it.

To explain how these quantities arise, let's consider the ridgeless case (i.e., $\lambda = 0$) and return to the geometric picture from Section 3.2. First of all, note that Λ is just the energy of the tail of the covariance, and as we already mentioned in Section 3.1, Λ is just implicit regularization that the tail imposes on the learning problem. Next, the term V corresponds to $\mathbb{E}_{\boldsymbol{y}}\|\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2$, which, as discussed in Section 3.1, is nothing but the variance term from the regression setting of Chapter 2. As in that chapter, V is bounded by a constant, but can be arbitrarily small. The quantity ΔV controls deviations of $\|\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2$ with respect to the randomness in $\boldsymbol{y}$.

The terms B and M arise as $B \approx \|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}}^2$ and $M \approx \boldsymbol{\mu}^\top\boldsymbol{\mu}_\perp$. Once again, one can notice that $\|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}}^2$ is exactly the bias term from Chapter 2. Interestingly, $\lozenge$, which is a rescaling of $\sqrt{B}$, also controls the magnitude of $\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y}$.

One may notice that our expressions for V and B are somewhat different from the main bounds in Chapter 2 (see Sections 2.4 and 2.7). This is because we choose a different presentation strategy: while Chapter 2 gives bounds in a simpler form, they are only tight for the right choice of k. The way we formulate the bounds in this chapter makes them tight for any choice of k under which the assumptions are satisfied, at the cost of worse-looking expressions. We elaborate more on differences in techniques with Chapter 2 in Section 3.5. A formal connection can be established by the following proposition:

Proposition 40.

$$\begin{aligned} V \leq& \frac{k}{n} + \Lambda^{-2} n \sum_{i>k} \lambda_i^2, \\ B \leq& n^{-2}\Lambda^2 \left\|\boldsymbol{\mu}_{0:k}\right\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}} + \left\|\boldsymbol{\mu}_{k:\infty}\right\|^2_{\boldsymbol{\Sigma}_{k:\infty}}. \end{aligned}$$

Proof.

$$\begin{aligned} V = n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2} n \sum_{i>k}\lambda_i^2 \leq& \\ \leq n^{-1}\mathrm{tr}\left(\boldsymbol{I}_k\right) + \Lambda^{-2} n \sum_{i>k}\lambda_i^2 = \frac{k}{n} + \Lambda^{-2} n \sum_{i>k}\lambda_i^2&, \end{aligned}$$

$$\begin{aligned} B = n^{-2}\Lambda^2 \left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + \left\|\boldsymbol{\mu}_{k:\infty}\right\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \leq& \\ \leq n^{-2}\Lambda^2\left\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + \left\|\boldsymbol{\mu}_{k:\infty}\right\|^2_{\boldsymbol{\Sigma}_{k:\infty}} = n^{-2}\Lambda^2 \left\|\boldsymbol{\mu}_{0:k}\right\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}} + \left\|\boldsymbol{\mu}_{k:\infty}\right\|^2_{\boldsymbol{\Sigma}_{k:\infty}}&. \end{aligned}$$

□

The fact that the bounds are the same up to a constant multiplier for the right choice of k can be seen from Lemma 47.

When it comes to noisy labels, informally, V also controls $\|\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}^2$ and $\lozenge$ controls $\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$. These statements are only informal because in our proof we don't deal with the vector $\tilde{\boldsymbol{y}}$ directly. It is more convenient to work with the vector $\Delta\boldsymbol{y} := \hat{\boldsymbol{y}} - \boldsymbol{y}$ as it has i.i.d. coordinates.

Finally, by virtue of algebra, our results extend to ridge regression, not just MNI. However, among the quantities defined above, only Λ has explicit dependence on the regularization parameter λ, all the other quantities depend on λ through Λ. Thus, explicit regularization simply adds to the implicit regularization from the data, without qualitatively changing the results.

When it comes to the interpretation from Section 3.1, it is not hard to see that N is within a constant factor of $\boldsymbol{\mu}^\top(\boldsymbol{\Sigma} + \Lambda n^{-1}\boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$, while $n\lozenge^2$ is within a constant factor of $\|(\boldsymbol{\Sigma} + \Lambda n^{-1}\boldsymbol{I}_p)^{-1}\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2$ (see Lemma 48 for a precise statement).

Some useful relations between those quantities are shown by the following lemma, whose proof can be found in Appendix B.3.

Lemma 41 (Relations between the main quantities)**.** *Suppose that*

$$k \leq n \quad \textit{and} \quad \Lambda > n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}. \tag{3.12}$$

Then

$$n\lozenge^2 \leq N, \quad n\lozenge^2 \leq N\sqrt{n\Delta V}, \quad V \leq 2, \quad \Delta V \leq \frac{3}{n}, \quad \Delta V \leq 4V.$$

Lower bound

Our main lower bound on the quantity $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ is given by the following theorem.

Theorem 42 (Main lower bound). *For any $c_B > 0, L > 1$ there exists a constant c that only depends on c_B and L, such that the following holds. Assume that $\eta < c^{-1}$, $k < n/c$, and*

$$\Lambda > cn\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}.$$

For any $t \in (0, \sqrt{n}/c)$, conditionally on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$, with probability at least $1 - ce^{-t^2/2}$ over the draw of $(\boldsymbol{y}, \hat{\boldsymbol{y}})$, the following inequalities hold for a certain scalar $S > 0$:

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} \geq c^{-1}N - ct\lozenge, \tag{3.13}$$

$$S\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}} \leq c\left([1 + N\sigma_\eta]\sqrt{V + t^2\Delta V} + \lozenge\sqrt{n}\right). \tag{3.14}$$

That is, if $N > 2c^2 t\lozenge$, then on the same event,

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} \geq \frac{1}{2c^2}\frac{N}{[1 + N\sigma_\eta]\sqrt{V + t^2\Delta V} + \lozenge\sqrt{n}}.$$

It is informative to explain how this result relates to the expressions we derived in Section 3.2. Recall that we restrict ourselves to the case $\lambda = 0$ in that section, that is, $\boldsymbol{w}_{\text{ridge}} = \boldsymbol{w}_{\text{MNI}}$. First of all, for $\lambda = 0$, $S = \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\tilde{\boldsymbol{w}}^c_{\text{MNI}}\|^2$. As in Section 3.2, let's start with the noiseless case, that is, $\eta = \sigma_\eta = 0$ and $\boldsymbol{y} = \hat{\boldsymbol{y}}$.

$$\begin{aligned} S\boldsymbol{w}^c_{\text{MNI}} =& \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\tilde{\boldsymbol{w}}^c_{\text{MNI}} \\ =& \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_\perp + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}. \end{aligned}$$

We need to bound $S\boldsymbol{\mu}^\top \boldsymbol{w}^c_{\text{MNI}}$ from below and $S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$ from above, so we write

$$\begin{aligned} S\boldsymbol{\mu}^\top \boldsymbol{w}^c_{\text{MNI}} =& \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}, \\ S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \leq& \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}} + (1 + |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|)\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}. \end{aligned}$$

The bound from Theorem 42 can now be obtained by plugging in the following:

$$\begin{aligned} \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \approx& n\Lambda^{-1}, \\ \|\boldsymbol{\mu}_\perp\|^2 \approx& M, \\ |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| \lesssim& t\lozenge, \\ \|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}} \lesssim& \sqrt{B}, \\ \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}} \lesssim& \sqrt{V + t\Delta V}. \end{aligned}$$

As was stated in Section 3.2, the contribution of the term $\boldsymbol{\mu}_\perp$ dominates the contribution of $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}, \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ in both those bounds, that is

$$n\Lambda^{-1}M \gtrsim t^2\Diamond^2, \quad n\Lambda^{-1}\sqrt{B} = \sqrt{n}\Diamond \gtrsim t\Diamond\sqrt{V + t\Delta V}.$$

Overall, we get the bounds

$$S\boldsymbol{\mu}^\top \boldsymbol{w}^c_{\text{MNI}} \gtrsim N - ct\Diamond, \quad S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \lesssim \sqrt{V + t\Delta V} + \sqrt{n}\Diamond.$$

Now let's consider the case with label-flipping noise. As we already mentioned in Section 3.2, due to certain algebraic cancellations the formula for the scalar product in the noisy case is very similar to the formula in the noiseless case:

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}} = \boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\|\boldsymbol{\mu}_\perp\|^2 + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}.$$

Since the vector $\hat{\boldsymbol{y}}$ is just a noisy version of the vector $\boldsymbol{y}$, the quantities $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}$ and $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}$ are close to $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ and $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ correspondingly, which yields the same bound on the scalar product as in the noiseless case.

When it comes to the denominator, as we already stated in Section 3.2, only one extra term survives compared to the noiseless case, the other terms get dominated. To show exactly how that happens, write

$$S\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \le S\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} + (\xi + |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}|)S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}.$$

We already have the upper bound on $S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$. The other bounds come from the following inequalities:

$$\begin{aligned} S =&(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 \\ \lesssim&(1 + t\Diamond)^2 + N, \\ \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} \lesssim&\sigma_\eta\sqrt{V + t\Delta V}, \\ \xi \approx&1, \\ |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}| \lesssim&\sigma_\eta t\Diamond. \end{aligned}$$

Combining everything together yields

$$S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \lesssim \left((1 + t\Diamond)^2 + N\right)\sigma_\eta\sqrt{V + t\Delta V} + (1 + \sigma_\eta t\Diamond)\left(\sqrt{V + t\Delta V} + \sqrt{n}\Diamond\right).$$

By Lemma 41 and since $t < \sqrt{n}$, N dominates $t^2\Diamond^2$. Moreover, $\sigma_\eta N\sqrt{V + t\Delta V}$ dominates $\sigma_\eta t\Diamond \cdot \sqrt{n}\Diamond$, Dropping those dominated terms leaves us with

$$S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \lesssim (3t\Diamond + N)\,\sigma_\eta\sqrt{V + t\Delta V} + \sqrt{V + t\Delta V} + \sqrt{n}\Diamond.$$

Finally, the term $3t\Diamond\sigma_\eta\sqrt{V + t\Delta V}$ is dominated by $\sqrt{n}\Diamond$, which gives us the final bound:

$$S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \lesssim (1 + N\sigma_\eta)\,\sqrt{V + t\Delta V} + \sqrt{n}\Diamond.$$

Recall, however, that this derivation is only informal, as our proof does not give rigorous bounds on quantities involving $\tilde{\boldsymbol{y}}$, as we deal with $\Delta\boldsymbol{y} = \hat{\boldsymbol{y}} - \boldsymbol{y}$ instead.

Upper bound

Even though a lower bound on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ is more important than the upper bound (since a lower bound provides a guarantee for high classification accuracy), obtaining an upper bound turns out to be more technically challenging. We elaborate on that in Section 3.5, where we explain what additional assumptions and tricks we needed to prove the upper bound and why.

Because of those technical difficulties, we only provide the upper bound for the regime without label-flipping noise. It is given by the following:

Theorem 43 (Main upper bound)**.** *Suppose that $\eta = 0$ — there is no label-flipping noise, and the rows of $\boldsymbol{Z}$ are σ_x-sub-Gaussian. For any $L > 1$ there are large constants a, c that only depend on σ_x and L and an absolute constant c_y such that the following holds. Suppose that $k < n/c$ and*

$$\Lambda > c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

Assume that $\boldsymbol{Q}_{k:\infty}$ is independent from $\boldsymbol{Q}_{0:k}$, and the distribution of $\boldsymbol{Q}_{k:\infty}$ is symmetric.

1. *If $N < a^{-1}\lozenge$ then with probability at least $c_y^{-1}(\mathbb{P}(\mathscr{A}_k(L)) - ce^{-n/c})_+$,*

$$\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} < 0.$$

Here u_+ denotes $u \vee 0$ for any $u \in \mathbb{R}$.

2. *If $N \geq a^{-1}\lozenge$ then for any $t \in (0, \sqrt{n}/c_y)$ the probability of the event*

$$\left\{\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} \leq c(1+t)\frac{N}{\sqrt{V + n\lozenge^2}}\right\}$$

is a least

$$(c_y^{-1} - c_y e^{-t^2/c_y} - c_y e^{-n/c})_+(\mathbb{P}(\mathscr{A}_k(L)) - ce^{-n/c})_+.$$

Tight bound for a quantile

Theorems 42 and 43 give lower and upper bounds on the quantity $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ correspondingly. That quantity, however, is random, and the bounds depend on the parameter t which controls the probability with which the bound holds. We don't expect those bounds to be sharp for all possible values of t, but we show that they are sharp for the case when t is a constant.

Definition 44. *For any $\varepsilon \in (0,1)$ we denote by α_ε the following ε-quantile of the distribution of $\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}/\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}$:*

$$\alpha_\varepsilon := \inf\left\{\alpha \in \mathbb{R} : \mathbb{P}\left(\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} < \alpha\right) > \varepsilon\right\}. \tag{3.15}$$

The following theorem shows that our upper and lower bounds on α_ε are within a constant factor of each other when ε is set to a certain absolute constant.

Theorem 45 (Tightness of the bounds)**.** *Suppose that the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. Suppose that $\eta = 0$ — there is no label-flipping noise. For any $L > 1$ there exist constants a, c that only depend on L, σ_x and absolute constants ε, δ such that the following holds. Suppose that $n > c$, $k < n/c$,*

$$\Lambda > c\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right),$$

and the probability of the event $\mathscr{A}_k(L)$ is at least $1-\delta$. Assume that $\boldsymbol{Q}_{k:\infty}$ is independent from $\boldsymbol{Q}_{0:k}$, and the distribution of $\boldsymbol{Q}_{k:\infty}$ is symmetric.

Then

$$\alpha_\varepsilon \le c\frac{N}{\sqrt{V} + \sqrt{n}\diamondsuit}.$$

If additionally $N \ge a\diamondsuit$, then

$$\alpha_\varepsilon \ge c^{-1}\frac{N}{\sqrt{V} + \sqrt{n}\diamondsuit}.$$

Regimes of the lower bound

In this section we discuss which form the bound from Theorem 42 can take depending on which terms dominate in the expressions.

First of all, the bound depends on the choice of t and η. We put t to be a large constant, and η to be a small constant. That is, the bound will hold with high constant probability, and the probability of label-flipping noise will be a small constant.

Let's think about the classification problem in the following way: treat the covariance $\boldsymbol{\Sigma}$, the number of data points n and the direction of $\boldsymbol{\mu}$ as fixed, and treat the magnitude of vector $\boldsymbol{\mu}$ as a parameter.

The bound on $S\boldsymbol{\mu}^\top\boldsymbol{w}_{\text{ridge}}$ from Equation (3.13) has two terms (up to constant multipliers): N and $-\diamondsuit$. The term N is quadratic in $\boldsymbol{\mu}$, while $\diamondsuit$ is linear. Thus, for small $\boldsymbol{\mu}$ the second term will dominate and the bound will be negative, but when $\boldsymbol{\mu}$ becomes large, N will dominate.

When it comes to the bound on $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ from Equation (3.14), it has three terms: $\sqrt{V}$, $\diamondsuit\sqrt{n}$, and $N\sqrt{V}$. The first term doesn't scale with $\boldsymbol{\mu}$, the second is linear in $\boldsymbol{\mu}$ and the third is quadratic. Thus, when $\boldsymbol{\mu}$ is small, the term V will dominate the bound, and when it is large, $N\sqrt{V}$ will dominate. Note that in the noiseless case ($\eta = 0$), there is no term $N\sqrt{V}$, and for large magnitude of $\boldsymbol{\mu}$ the term $\diamondsuit\sqrt{n}$ will dominate in the bound.

Now, let's investigate how those transitions relate to each other. According to Lemma 41, $\sqrt{n}\diamondsuit^2 \le N\sqrt{\Delta V} \le 2N\sqrt{V}$, which implies

$$\frac{\sqrt{V}}{\diamondsuit\sqrt{n}} \ge \frac{\diamondsuit}{2N}.$$

That is, if $\Diamond$ is at least a constant times N, then $\sqrt{V}$ is at least a constant times $\Diamond\sqrt{n}$. Moreover,

$$N\sqrt{V} \le \frac{N^2}{n\Diamond^2}\sqrt{V},$$

so if $\Diamond$ dominates N, then $\sqrt{V}$ dominates $N\sqrt{V}$ (and even $nN\sqrt{V}$). Thus we see that the transition in the bound for $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ happens first: N starts dominating $\Diamond$ earlier than $\sqrt{V}$ stops dominating $\Diamond\sqrt{n}$ and $N\sqrt{V}$.

Another question is whether for a constant value of η, the term $\Diamond\sqrt{n}$ can dominate in the bound on $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$. As turns out, it may or may not happen depending on the relation between $\boldsymbol{\Sigma}$ and the direction of $\boldsymbol{\mu}$. To see this, consider two examples: first, assume that $\boldsymbol{\mu} = m\boldsymbol{e}_1$ — a vector supported on the first coordinate. For such choice of $\boldsymbol{\mu}$ we have

$$\Diamond\sqrt{n} = \frac{\sqrt{\lambda_1}m}{\lambda_1 + \Lambda/n}, \quad N = \frac{m^2}{\lambda_1 + \Lambda/n}.$$

We see that if $\lambda_1 \gg \Lambda/n$ and $1 \gg V$, then for $m = \sqrt{\lambda_1}$ we will get $\Diamond\sqrt{n} \gg (1+N)\sqrt{V}$.

On the other hand, consider $\boldsymbol{\mu} = m\boldsymbol{e}_p$ — a vector supported on the last coordinate. Then

$$\Diamond\sqrt{n} = n\Lambda^{-1}\sqrt{\lambda_p}m, \quad N = n\Lambda^{-1}m^2.$$

If it so happens that $V \gg n\lambda_p/\Lambda$ (which is possible as λ_p can be arbitrarily small), then we can write

$$(1+N)\sqrt{V} = (1+n\Lambda^{-1}m^2)\sqrt{V} \gg (1+n\Lambda^{-1}m^2)\sqrt{n\Lambda^{-1}\lambda_p} \ge 2\sqrt{n\Lambda^{-1}m^2}\sqrt{n\Lambda^{-1}\lambda_p} = 2\Diamond\sqrt{n}.$$

That is, for such choice of covariance and $\boldsymbol{\mu}$, the term $\Diamond\sqrt{n}$ cannot dominate in the bound for $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ for any choice of m.

Table 3.1 summarizes the discussion of the regimes we had so far.

Magnitude of $\boldsymbol{\mu}$	small	medium	large	very large
Bound on $\boldsymbol{\mu}^\top\boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$	$-\Diamond/\sqrt{V}$	$N/\sqrt{V}$	$N/(\sqrt{n}\Diamond)$	$1/\sqrt{V}$
Occurs in noiseless case $\eta = 0$	yes	yes	yes	no
Occurs in noisy case $\eta = c^{-1}$	yes	yes	sometimes	yes

Table 3.1: Regimes of the main bound

Finally, recall the transition that happens in the structure of the noisy solution, namely in Equation (3.7) from Section 3.2, which was derived for the MNI solution with label-flipping noise:

$$\boldsymbol{w}_{\text{MNI}} = \boldsymbol{Q}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}} + (\xi - \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}})\boldsymbol{w}_{\text{MNI}}^c.$$

We see that the clean solution is multiplied by a scalar $\xi - \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$. As discussed in Section 3.2, $\xi \approx 1-2\eta \approx 1$, while $\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ is a random variable with symmetric distribution, which is also linear in $\boldsymbol{\mu}$. As the magnitude of $\boldsymbol{\mu}$ grows, the magnitude of the zero-mean term $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ also grows, and at some point it starts dominating the term ξ. At that point the contribution of the clean solution to the noisy solution becomes "washed out", and the component of $\boldsymbol{w}_{\text{MNI}}$ in the direction of $\boldsymbol{w}^c_{\text{MNI}}$ becomes close to zero mean instead of being positive (recall $\tilde{\boldsymbol{y}} \perp \boldsymbol{w}^c_{\text{MNI}}$).

Let's investigate how this qualitative transition in the structure of the noisy solution is related to other transitions that we discussed previously. Unfortunately, as we already mentioned before, our arguments don't provide a rigorous bound on $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$. Because of that we can only speculate about the magnitude of $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$, and here we will assume that for constant noise level we have $|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}| \approx |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|$, with the magnitude of the latter term being controlled by $\diamondsuit$. Thus, the transition happens when $\diamondsuit$ becomes larger than 1. In this case, however, we can write using Lemma 41

$$\diamondsuit \gtrsim 1,\ N \geq n\diamondsuit^2 \gtrsim n \gg 1,\ N\sqrt{V} \gg \sqrt{V},\ N\sqrt{V} \geq N\sqrt{\Delta V}/2 \geq \sqrt{n}\diamondsuit^2 \gtrsim \sqrt{n}\diamondsuit.$$

Thus, in this regime, the term $N\sqrt{V}$ is the largest (up to a constant multiplier) in the bound on $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$.

To conclude, let's tie the regimes we discussed in this section to the geometrical pictures of Section 3.2.

For the case without label-flipping noise, Equation (3.6) shows that the vector $\boldsymbol{w}_{\text{MNI}}$ is proportional to

$$\frac{\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}} + \boldsymbol{\mu}_\perp + \frac{\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}.$$

The term $\boldsymbol{\mu}_\perp$ dominates the last term, and acts effectively as $(n\Lambda^{-1}\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$. As a result, as stated in Section 3.1, $\boldsymbol{w}_{\text{MNI}}$ performs as $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$. When $\boldsymbol{\mu}$ is small in magnitude, and $\diamondsuit$ dominates N, the "noise term" $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ will dominate in terms of both $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}$ and $\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$, and the quantity $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}$, resulting in a negative bound. As the magnitude of $\boldsymbol{\mu}$ grows, N will become larger than $\diamondsuit$, but V will still dominate over $\sqrt{n}\diamondsuit$, so the term $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ dominates in terms of the scalar product with $\boldsymbol{\mu}$, while the term $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ will still dominate in terms of the norm in $\boldsymbol{\Sigma}$. The bound in this regime is $N/\sqrt{V}$ up to a constant factor. Finally, as $\boldsymbol{\mu}$ becomes even larger, $\boldsymbol{w}_{\text{MNI}}$ performs effectively as $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$, and the bound becomes $N/(\sqrt{n}\diamondsuit)$ up to a constant factor.

The noisy case is harder to explain without an intuitive explanation why the formula for $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}$ is so similar to the formula for $\boldsymbol{\mu}^\top \boldsymbol{w}^c_{\text{MNI}}$ (end of Section 3.2). Nevertheless, that similarity yields that the condition when the bound becomes positive is the same: $\boldsymbol{\mu}$ should be large enough so that N dominates $\diamondsuit$. After that, the bound for the noisy case goes through the same regimes: from $N/\sqrt{V}$ to $N/(\sqrt{n}\diamondsuit)$, and then through a new regime: $1/\sqrt{V}$. The regime $N/(\sqrt{n}\diamondsuit)$ may or may not appear, depending on how the energy of $\boldsymbol{\mu}$ is distributed across the eigendirections of $\boldsymbol{\Sigma}$.

Now let's explain geometrically how those transitions happen. Equation (3.7) from Section 3.2 states

$$\boldsymbol{w}_{\text{MNI}} = \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}} + (\xi - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}})\boldsymbol{w}^c_{\text{MNI}}.$$

In that section we also explained that the noisy solution is very similar to the noiseless solution in terms of scalar product with $\boldsymbol{\mu}$, while it's norm in $\boldsymbol{\Sigma}$ only picks up one additional term: the norm of $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$. That is

$$\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}} \approx \boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}^c, \quad \|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \approx \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{w}_{\text{MNI}}^c\|_{\boldsymbol{\Sigma}}.$$

For small $\boldsymbol{\mu}$ and constant η we have

$$\|\boldsymbol{w}_{\text{MNI}}^c\|_{\boldsymbol{\Sigma}} \approx \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}} \approx \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} \approx \sqrt{V}.$$

That is, for small $\boldsymbol{\mu}$ not only the scalar product in $\boldsymbol{\mu}$, but also the norm in $\boldsymbol{\Sigma}$ is similar for the noisy and the noiseless solution. Thus, the first regime that the bound goes through is the same as the initial regime for the clean solution, with the bound $N/\sqrt{V}$. As $\boldsymbol{\mu}$ grows further, $\|\boldsymbol{w}_{\text{MNI}}^c\|_{\boldsymbol{\Sigma}}$ may grow and start dominating over $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}$, so the bound may go over the second regime of the clean solution: $N/(\sqrt{n}\Diamond)$. Eventually, however, $\boldsymbol{w}_{\text{MNI}}^c$ converges to zero as $\boldsymbol{\mu}$ goes to infinity, so $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}$ dominates $\|\boldsymbol{w}_{\text{MNI}}^c\|_{\boldsymbol{\Sigma}}$. So it becomes

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}}{\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}}} \approx \frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}^c}{\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}}.$$

Recall that we are in the regime when $\boldsymbol{\mu}$ is large, and $\boldsymbol{w}_{\text{MNI}}^c$ has high classification accuracy. Recall also that $\boldsymbol{w}_{\text{MNI}}^c$ is defined as an interpolator of labels ±1, and $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}^c$ is the label that it assigns to the center of the positive cluster. Thus, $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}^c \approx 1$. Since for constant η we have $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} \approx \sqrt{V}$, we get the final regime, when the bound becomes $1/\sqrt{V}$. Note that this quantity loses dependence on $\boldsymbol{\mu}$ completely, and the only condition that makes the bound large is that V has to be small. In Section 3.1, as well as the beginning of Section 3.4, we've seen that V is (up to a multiplicative constant) equivalent to the bound on the variance term obtained by Chapter 2. Because of that, the sufficient condition for benign overfitting that we get in the large $\boldsymbol{\mu}$ regime is equivalent to the benign overfitting condition in linear regression.

Benign overfitting

One interesting phenomenon that Theorem 42 implies is that the misclassification error can be arbitrarily close to zero even if we have a small constant level of label-flipping noise. One can see that it happens for $t \ll \sqrt{n}$, $V \ll 1$, and $N \gg \sqrt{V} + t/\sqrt{n} + \Diamond\sqrt{n}$ (which is a form of an SNR condition). In order to simplify the presentation, we formulate a rigorous corollary for Gaussian distribution:

Corollary 46. *Suppose the rows of matrix $\boldsymbol{Q}$ come as i.i.d. samples from a Gaussian distribution. Take $\lambda = 0$ (that is, $\boldsymbol{w}_{ridge}$ coincides with $\boldsymbol{w}_{MNI}$). There exists a large absolute constant c such that the following holds for any $C > 1$.*

Assume all the following:

the noise is bounded by a constant: $$\eta < c^{-1}, \tag{3.16}$$

the spiked part of the covariance has low dimension: $$k < n/(cC^2), \tag{3.17}$$

the tail of the covariance has high effective rank: $$\sum_{i>k} \lambda_i > cn\lambda_{k+1} \vee cC\sqrt{n\sum_{i>k}\lambda_i^2}, \tag{3.18}$$

the scale of $\boldsymbol{\mu}$ *is large enough:* $$N \geq 1 + cC\left(\sqrt{V} + \Diamond\sqrt{n}\right). \tag{3.19}$$

Then with probability at least $1 - ce^{-n/(cC)^2}$*,*

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{MNI}}{\|\boldsymbol{w}_{MNI}\|_{\boldsymbol{\Sigma}}} \geq C. \tag{3.20}$$

Proof. First of all, due to Equation (3.18), if c is large enough, by Lemma 37, for absolute constants c_A, L the probability of the event $\mathscr{A}_k(L)$ is at least $1 - c_A e^{-n/c_A}$. Moreover, due to Equation (3.17) and Lemma 39, for an absolute constant c_B the event $\mathscr{B}_k(c_B)$ holds with probability at least $1 - c_B e^{-n/c_B}$.

Next, by Theorem 42, for an absolute constant c_1 on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$ we have for a certain scalar $S > 0$ for any $t \in (0, \sqrt{n}/c_1)$, with probability at least $1 - c_1 e^{-t^2/2}$,

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{\mathrm{MNI}} \geq c_1^{-1}N - c_1 t\Diamond, \tag{3.21}$$

$$S\|\boldsymbol{w}_{\mathrm{MNI}}\|_{\boldsymbol{\Sigma}} \leq c_1\left([1 + N\sigma_\eta]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n}\right). \tag{3.22}$$

Since $t < \sqrt{n}/c_1$, if c is large enough, due to Equation (3.19), for an absolute constant c_2 Equation (3.21) implies $S\boldsymbol{\mu}^\top \boldsymbol{w}_{\mathrm{MNI}} \geq c_2^{-1}N$.

So far, for an absolute constant c_3 we have with probability at least $1 - c_1 e^{-t^2/2}$,

$$\begin{aligned}\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{\mathrm{MNI}}}{\|\boldsymbol{w}_{\mathrm{MNI}}\|_{\boldsymbol{\Sigma}}} &\geq c_3^{-1}\frac{N}{[1 + N\sigma_\eta]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n}} \\ &\geq \frac{1}{2c_3}\left(\frac{N}{\Diamond\sqrt{n} + \sqrt{V + t^2\Delta V}} \wedge \frac{1}{\sigma_\eta\sqrt{V + t^2\Delta V}}\right).\end{aligned}$$

For simplicity, we use Lemma 41 to bound $\Delta V < 3/n$. We see that to achieve the bound (3.20), we can show the following two conditions:

$$N \geq 2Cc_3\left(\Diamond\sqrt{n} + \sqrt{V} + \sqrt{3t^2/n}\right), \quad \sigma_\eta(\sqrt{V} + \sqrt{3t^2/n}) \leq (2c_3C)^{-1}.$$

Recall that $\eta < c^{-1}$, so $\sigma_\eta < 1$. The two conditions above can be achieved by first taking $t^2 = n/(c_4C^2)$ for a large enough absolute constant c_4. Then, Equation 3.19 implies the first condition, while Equations (3.17) and (3.18) imply the second due to Proposition 40. □

We believe these sufficient conditions for benign overfitting in classification to be novel. For example, in a recent paper, [58] obtain similar conditions only for the case of isotropic data (i.e., $\boldsymbol{\Sigma} = \boldsymbol{I}_p$); see their Theorem 7.

3.5 Proof outline and different forms of the main quantities

Lower bound proof sketch

On a high level we follow the same logic as in Section 2.6 of Chapter 2: use algebraic formulas for the solution to disentangle the contribution of components $0:k$ from the components $k:\infty$, then plug in concentration inequalities. It is, however, a more difficult task in the case of the mixture model because clusters are not zero-mean, and have centers $\pm\boldsymbol{\mu}$. Algebraically, we have $\boldsymbol{X} = \boldsymbol{Q} + \boldsymbol{y}\boldsymbol{\mu}^\top$, and the rank-one correction $\boldsymbol{y}\boldsymbol{\mu}^\top$ prohibits straightforward application of the machinery from Chapter 2. Thus, the first step in the proof is to transform the expression for $\boldsymbol{w}_{\text{ridge}} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I}_n)^{-1}\hat{\boldsymbol{y}}$ into a form that operates with the inverse of $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{Q}^\top + \lambda\boldsymbol{I}_n$ instead of $\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I}_n$. The corresponding formula is obtained in Appendix B.1 Lemma 89.

After obtaining the formula in Lemma 89, we derive sharp bounds on all the terms that appear in it. Note that we have two independent sources of randomness in our setting: randomness from the matrix $\boldsymbol{Q}$ and randomness from the labels $(\boldsymbol{y}, \hat{\boldsymbol{y}})$. We start by addressing the second source, making the high probability statements over the draw of $(\boldsymbol{y}, \hat{\boldsymbol{y}})$ conditionally on $\boldsymbol{Q}$ in Lemma 98, Appendix B.4.

As a result of Lemma 98 we reduce the problem to bounding expressions that only depend on $\boldsymbol{Q}$. At this point we can apply the ideas developed in Chapter 2 to bound those quantities, which we do with some modifications. As in Section 2.6, we start by deriving algebraic bounds that hold almost surely on the event that the matrix $\boldsymbol{A}_k$ is PD. Those bounds can be found in Lemma 99, Appendix B.5. Some of the terms that we bound there have already appeared in Section 2.6, namely $\|\boldsymbol{\mu}_{\underset{\sim}{\perp}}\|_{\boldsymbol{\Sigma}}^2$ and $\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})$ are exactly bias and variance of the regression problem. We could, in principle, reuse the results of that chapter, but we do a new slightly different derivation to obtain them in a different form. Namely, we directly use the Sherman-Morrison-Woodbury (SMW) identity (Lemma 100), the most important embodiment of which is $\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k} = \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}$, to obtain our algebraic decompositions. After that, we use Lemma 101 to substitute $\boldsymbol{\Sigma}_{0:k}^{1/2}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{1/2}$ by $\alpha^{-1}(\beta^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-1}$, where α, β are scalars that concentrate within a constant factor of their typical value. The alternative strategy from Section 2.6 would be to say that the matrix $\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}$ is dominated by $\boldsymbol{I}_k$ for $k = k^*$. Because of that, our results are sharp for any choice of k and we obtain upper bounds in the same form as lower bounds right away, while in Chapter 2 we obtained them in different forms and needed to do a separate conversion to show that they coincide for $k = k^*$. We pay for that, however, with a bulkier form of the bounds in this chapter.

The bounds from Lemma 99 are formulated in terms of quantities that we assume to be concentrated around their typical values on the events $\mathscr{A}_k(L)$ and $\mathscr{B}_k(c_B)$. Plugging those values into the bounds is done in Lemma 102, Appenxix B.6. Finally, we finish the proof of Theorem 42 in Appendix B.7: first, we combine the bounds from Lemmas 98 and 102 in

Lemma 103. Then we plug the result into the formulas from Lemma 89.

Upper bound proof sketch

When we deal with the bounds within a constant factor, it is usually more difficult to obtain a bound from below than from above. This is because to bound a sum from above one can use a triangle inequality $|a+b| \le |a| + |b|$ to reduce the problem to bounding separate terms from above. If, however, we want to bound a sum $|a+b|$ from below, the triangle inequality yields $|a+b| \ge (|a| - |b|) \vee (|b| - |a|)$, which is only sharp when one term dominates another in magnitude.

In our case, we want to bound the fraction $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$. To bound it from below we bound $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ from below and $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ from above. Moreover, it turns out that there is only one term in the expression for $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ that we actually need to bound from below, other terms play the role of noise and can be bounded from above in absolute value. Because of that, bounding $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ from below is a much more straightforward task than bounding it from above.

Two problems arise when we bound $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ from above: first, we need to bound $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ from below, and there is no one dominating term in its expression. Second problem is to show that the numerator $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ will be negative with constant probability if N is not large enough compared to $\diamondsuit$, for which we also need to bound the magnitude of the noise terms in the numerator from below.

To alleviate the problem with the triangle inequality we resort to the following trick: we assume that $\boldsymbol{Q}_{k:\infty}$ has a symmetric distribution and is independent from $\boldsymbol{Q}_{0:k}$. This means that the joint distribution of $(\boldsymbol{Q}_{0:k}, \boldsymbol{Q}_{k:\infty}, \boldsymbol{y})$ is the same as that of $(\boldsymbol{Q}_{0:k}, \varepsilon_q \boldsymbol{Q}_{k:\infty}, \varepsilon_y \boldsymbol{y})$, where we introduced two new Rademacher random variables $(\varepsilon_q, \varepsilon_y)$, which are independent of all previously defined random variables and from each other. The basic idea behind the introduction of these random variables is as follows: suppose ε is a Rademacher random variable, which is independent of random variables a, b. Then, conditionally on a, b, with probability 0.5 over the draw of ε, $|a + \varepsilon b| = |a| + |b|$. If we now have high-probability lower bounds on $|a|$ and $|b|$, then we get a constant probability lower bound on $|a+b|$.

To explain how this idea applies to the quantities that we need to bound, we will need to look at their exact expressions. Denote

$$\bar{\boldsymbol{Q}} := [\boldsymbol{Q}_{0:k}, \varepsilon_q \boldsymbol{Q}_{k:\infty}], \quad \bar{\boldsymbol{y}} := \varepsilon_y \boldsymbol{y}, \quad \bar{\boldsymbol{w}}_{\text{ridge}} := (\bar{\boldsymbol{Q}} + \bar{\boldsymbol{y}}\boldsymbol{\mu}^\top)^\top (\underbrace{\bar{\boldsymbol{Q}}\bar{\boldsymbol{Q}}^\top + \lambda \boldsymbol{I}_n}_{=\boldsymbol{A}})^{-1} \bar{\boldsymbol{y}}.$$

(Recall that for the upper bounds we only consider the case with no label-flipping noise, i.e., $\boldsymbol{y} = \hat{\boldsymbol{y}}$.) The expression for the numerator is now

$$\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}} = \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}} \boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}} + (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}}) \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}},$$

where $\bar{\boldsymbol{\nu}} = \bar{\boldsymbol{Q}}\boldsymbol{\mu}$, $\bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}} = (\boldsymbol{I}_p - \bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{Q}})\boldsymbol{\mu}$, and $\bar{S}$ is a scalar, which is non-negative with high probability. The proof of Theorem 42 already provides sharp high-probability bound on the term $\bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}}$, as well as the upper bound on $|\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}|$. Thus, the difficulty is only in proving a lower bound on $|\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}|$ to say that the term $\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}$ will make the numerator negative with constant probability unless N is large compared to $\Diamond$. The random variable ε_q helps as follows: with probability 0.5 over the draw of ε_q,

$$\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\|^2 = \|\boldsymbol{A}^{-1}(\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k} + \varepsilon_q \boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty})\|^2 \geq \|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\|^2 + \|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2.$$

We bound the terms $\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\|^2$ and $\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2$ from below in Lemma 104. We then use them in Lemma 105 to get the full upper bound on the numerator $\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}$.

When it comes to the denominator $\|\bar{S}\bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, the expression that we use is

$$\bar{S}\bar{\boldsymbol{w}}_{\text{ridge}} = (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}}.$$

The term $|\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{y}|\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{y}\|_{\boldsymbol{\Sigma}}$ is dominated by others, so we can reuse an upper bound on it from the proof of Theorem 42. When it comes to the remaining terms, note that $\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{y} = \varepsilon_y \bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$, while $\bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}} = \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}}$, which does not depend on ε_y. Thus, with probability 0.5 over ε_y, the cross-term that arises from those two terms is non-negative and can be ignored for the purposes of obtaining a lower bound. Next, note that $\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}} = \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}$ does not depend on ε_q, thus with probability 0.5 over the draw ε_q we can ignore the cross terms that arise from interaction between components $0 : k$ and $k : \infty$ in $\bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}}$. Overall, with probability at least 0.25 over the draw of $(\varepsilon_q, \varepsilon_y)$ we can ignore a few terms in the expression for $\|\bar{S}\bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ to obtain a lower bound on it. The precise statement is given in Lemma 106.

The strategy for the remainder of the proof of the upper bound is the same as in the proof of the lower bound: in Lemma 107 we make high-probability statements with respect to the draw of $\boldsymbol{y}$, and in Lemma 108 we make high probability statements with respect to the draw of $\boldsymbol{Q}$. We put together the lower bound on the denominator in Lemma 109 and combine it with the upper bound on the numerator in Theorem 43, whose proof is given in Appendix B.8.

Note that due to the nature of the proof, we only obtain the upper bounds with constant probability.

3.6 Effect of regularization

In this section we discuss the effects of the regularization on the accuracy on the learned classifier in the noiseless setting (i.e., $\eta = 0$). We will touch on the noisy setting in Section 3.6, but we can only talk about the dependence of the lower bound on regularization there since we don't provide a matching upper bound.

The main result that we have for the noiseless case is Theorem 45, which proves tightness of the bound for a quantile α_ε. Throughout this section we will study how changing regularization affects that quantile.

Before we start, however, let's introduce two alternative forms of the bound on that quantile, that are somewhat more useful in terms of tracking the effect of λ. As we already pointed out, these bounds are closely related to the bounds for the regression problem studied in Chapter 2, but have somewhat different form. In the next two lemmas we show how definitions of our quantities of interest could be alternatively defined to have similar form to the quantities from that chapter.

The following lemma gives a form of the bounds using the notion of k^* introduced in Section 2.3. This form corresponds to the form of the main results obtained in Section 2.4.

Lemma 47 (Bounds via k^*)**.** *Suppose that*

$$k \le n/2 \quad \textit{and} \quad \Lambda > n\lambda_{k+1}.$$

Define

$$\begin{aligned}
k^* :=& \min\left\{\kappa \in \{0, 1, \dots, k\} : \lambda + \sum_{i>\kappa} \lambda_i \ge n\lambda_{\kappa+1}\right\},\\
\Lambda_* :=& \lambda + \sum_{i>k^*} \lambda_i,\\
V_* :=& \frac{k^*}{n} + \Lambda_*^{-2} n \sum_{i>k^*} \lambda_i^2,\\
\diamondsuit_*^2 :=& n^{-1}\left\|\boldsymbol{\mu}_{0:k^*}\right\|^2_{\boldsymbol{\Sigma}_{0:k^*}^{-1}} + n\Lambda_*^{-2}\|\boldsymbol{\mu}_{k^*:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}},\\
N_* :=& \|\boldsymbol{\mu}_{0:k^*}\|^2_{\boldsymbol{\Sigma}_{0:k^*}^{-1}} + n\Lambda_*^{-1}\|\boldsymbol{\mu}_{k^*:\infty}\|^2.
\end{aligned}$$

Then

$$2N_* \ge N \ge N_*/2, \quad 2\diamondsuit_* \ge \diamondsuit \ge \diamondsuit_*/2, \quad 4V_* \ge V \ge V_*/4, \quad \Lambda_* \ge \Lambda \ge \Lambda_*/2.$$

The next lemma gives an alternative form of the bounds, which makes their dependence on k less pronounced. They are analogous to the bounds from Section 2.7.

Lemma 48 (Alternative form of the bounds)**.** *Suppose that $k < n$ and $\Lambda > n\lambda_{k+1}$. Denote*

$$N_a := \sum_i \frac{\mu_i^2}{\lambda_i + \Lambda/n}, \quad V_a := \sum_i \frac{\lambda_i^2/n}{(\lambda_i + \Lambda/n)^2}, \quad {\diamondsuit_a}^2 := \sum_i \frac{\lambda_i \mu_i^2/n}{(\lambda_i + \Lambda/n)^2}.$$

Then

$$N \ge N_a \ge N/2, \quad V \ge V_a \ge V/4, \quad \diamondsuit^2 \ge {\diamondsuit_a}^2 \ge \diamondsuit^2/4.$$

Note that this form of the results already appeared in our discussion in Section 3.1.

Now we are in position to return to studying the effect of regularization. To track changing values of regularization parameter λ, for the rest of this section we add it explicitly to the notation, i.e., in this section we will write $\alpha_\varepsilon(\lambda), \Lambda(\lambda), N(\lambda), V(\lambda), \diamondsuit(\lambda), \mathscr{A}_k(L, \lambda)$ etc. Note that if $L > 1$ and $\lambda' > \lambda$, then $\mathscr{A}_k(L, \lambda) \subseteq \mathscr{A}_k(L, \lambda')$, that is, we always only need to assume that $\mathscr{A}_k(L, \lambda)$ holds for the smallest value of the regularization parameter that we consider.

Increasing regularization never helps in the noiseless case

Due to Theorem 45, the main quantity of interest in the setting without label-flipping noise is $\frac{N(\lambda)}{\sqrt{V(\lambda)}+\sqrt{n}\lozenge(\lambda)}$. According to Lemma 48, this quantity is within a constant factor of $\frac{N_a(\lambda)}{\sqrt{V_a(\lambda)}+\sqrt{n}\lozenge_a(\lambda)}$. The first interesting observation is that $N_a(\lambda)/\lozenge_a(\lambda)$ is a non-increasing function of λ. To see this denote

$$t := \Lambda/n, \quad \boldsymbol{v} := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}, \quad \boldsymbol{w} := (\boldsymbol{\Sigma} + t\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mu} = (\boldsymbol{I}_p + t\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{v}.$$

With this notation, it becomes

$$\frac{N_a(\Lambda)}{\sqrt{n}\lozenge_a(\Lambda)} = \frac{\boldsymbol{v}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|},$$

which is non-increasing by the following lemma, whose proof can be found in Appendix B.9.

Lemma 49. *Consider a non-zero vector $\boldsymbol{v} \in \mathbb{R}^p$ and a PD symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{p\times p}$. Introduce the function $f : \mathbb{R}^p \to \mathbb{R}$ as $f(\boldsymbol{w}) = \boldsymbol{v}^\top\boldsymbol{w}/\|\boldsymbol{w}\|$. Then $f\left((\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{v}\right)$ is a non-increasing function of t on $[0, +\infty)$.*

This observation already suggests that increasing regularization should not lead to an increase in the bound. The only way that it could happen is when the term $\sqrt{V(\lambda)}$ dominates $\sqrt{n}\lozenge$ in the denominator. As it turns out, however, in this case the vector $\boldsymbol{\mu}$ cannot be large enough for the bound to be larger than a constant. A formal statement is given by the following lemma, which is proven in Appendix B.9.

Lemma 50 (Increasing the regularization cannot make the bound large)**.** *Suppose that $k < n$ and $\Lambda(\lambda) > n\lambda_k$. Then for some absolute constant $c > 0$ and any $\lambda' > \lambda$*

$$\frac{N(\lambda')}{\sqrt{V(\lambda')} + \sqrt{n}\lozenge(\lambda')} \le c\left(1 + \frac{N(\lambda)}{\sqrt{V(\lambda)} + \sqrt{n}\lozenge(\lambda)}\right).$$

Combining this with Theorem 45 gives the following.

Corollary 51. *Suppose that the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. For any $L > 1$ there exist constants a, c that only depend on L and σ_x and absolute constants δ, ε such that the following holds. Assume that $n > c$, $k < n/c$, $\mathbb{P}(\mathscr{A}_k(L, \lambda)) > 1 - \delta$, $N(\lambda) \ge a\lozenge(\lambda)$, and*

$$\Lambda(\lambda) > c\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

Suppose that $\boldsymbol{Q}_{k:\infty}$ has a symmetric distribution and is independent from $\boldsymbol{Q}_{0:k}$.

For every $\lambda' \ge \lambda$

$$\alpha_\varepsilon(\lambda') \le c\left(1 + \alpha_\varepsilon(\lambda)\right).$$

Proof. Take a, δ, ε the same as in Theorem 45. Denote the constant c from that theorem as c_1. Note that $\mathbb{P}(\mathscr{A}_k(L, \lambda')) \geq \mathbb{P}(\mathscr{A}_k(L, \lambda)) > 1 - \delta$, which means that Theorem 45 applies for all values of the regularization parameter $\lambda' > \lambda$. Thus,

$$\alpha_\varepsilon(\lambda') \leq c_1 \frac{N(\lambda')}{\sqrt{V(\lambda')} + \sqrt{n}\lozenge(\lambda')}, \quad \alpha_\varepsilon(\lambda) \geq c_1^{-1} \frac{N(\lambda)}{\sqrt{V(\lambda)} + \sqrt{n}\lozenge(\lambda)}.$$

Combining it with Lemma 50 and taking c large enough depending on c_1 yields the result. □

Note, however, that our argument only works if the probability of the event $\mathscr{A}_k(L, \lambda)$ is high for some constant L, and that $\Lambda(\lambda)$ is large compared to $n\lambda_{k+1}$. Increasing λ increases both $\Lambda(\lambda)$ and the probability of $\mathscr{A}_k(L, \lambda)$. Therefore, the results above don't say that smaller values of regularization are always better. A more precise interpretation would be "if λ is large enough so that $\Lambda(\lambda) \gg n\lambda_{k+1}$ and $\mathscr{A}_k(L, \lambda)$ holds with high probability, then there is no benefit from increasing it further".

Increasing regularization does nothing in some regimes

Even though we showed that there is not much use (in a certain sense) in increasing regularization, we haven't yet shown that it is harmful. For example, the following question arises: can decreasing regularization increase $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}(\lambda)/\|\boldsymbol{w}_{\text{ridge}}(\lambda)\|_{\boldsymbol{\Sigma}}$ by more than a constant factor? As it turns out, it depends on how $\boldsymbol{\mu}$ is spread across the eigendirections of $\boldsymbol{\Sigma}$. For example, increasing regularization will always preserve the bound within a constant factor if $\boldsymbol{\mu}$ is supported on the tail of the covariance or $\boldsymbol{\mu}$ is an eigenvector of the covariance, and $\boldsymbol{\mu}$ is large enough so that $V(\lambda)$ is dominated by $n\lozenge(\lambda)^2$. The formal statement is given by the following corollary, which is proven in Appendix B.9.

Corollary 52 (Regularization doesn't matter for certain $\boldsymbol{\mu}$)**.** *Suppose that the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. For any $L > 1$ there exist constants a, c that only depend on L, σ_x and absolute constants ε, δ such that the following holds. Suppose that $n > c$, $k < n/c$, $\mathbb{P}(\mathscr{A}_k(L, \lambda)) > 1 - \delta$, $N(\lambda) \geq a\lozenge(\lambda)$, and*

$$\Lambda(\lambda) > c\left(n\lambda_{k+1} \vee \sqrt{n \sum_{i>k} \lambda_i^2}\right).$$

Suppose that $\boldsymbol{Q}_{k:\infty}$ has a symmetric distribution and is independent from $\boldsymbol{Q}_{0:k}$.

If either for some $i \leq k$

$$\boldsymbol{\mu} = \mu_i \boldsymbol{e}_i, \quad \text{and} \quad \frac{n\lambda_i \mu_i^2}{(1 + n\lambda_i/\Lambda(\lambda))^2} \geq \sum_i \lambda_i^2,$$

(here $\boldsymbol{e}_i$ is the i-th eigenvector of $\boldsymbol{\Sigma}$), or

$$\|\boldsymbol{\mu}_{0:k}\| = 0 \quad \textit{and} \quad \sum_i \lambda_i^2 \leq n\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}},$$

then for any $\lambda' \geq \lambda$,

$$\alpha_\varepsilon(\lambda')/c \leq \alpha_\varepsilon(\lambda) \leq c\alpha_\varepsilon(\lambda').$$

The results that we obtained so far seem to contradict the conclusion made by [58], who considered a particular case of our model with Gaussian data and $k = 0$ and concluded that increasing regularization always decreases the classification error (see their Section 6.1), and checked that empirically in simulations. According to our results, increasing λ does not change $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}(\lambda)/\|\boldsymbol{w}_{\text{ridge}}(\lambda)\|_{\boldsymbol{\Sigma}}$ by more than a constant factor in this regime. There is no actual contradiction, because [58] only proved that their bound is decreasing. They neither proved that the bound is sharp, nor that it can decrease by more than a constant factor. We provide a detailed comparison with their results in Section 3.7.

Increasing regularization may cause harm by breaking the balance between the tail and the spiked part

Now let's investigate for which $\boldsymbol{\mu}$ having regularization as small as possible actually provides more than a constant factor gain. Lemma 47 gives, perhaps, the most convenient formulas to look at. For simplicity let's restrict ourselves to the case where $\boldsymbol{\mu}$ is large enough, so the term $\sqrt{V(\lambda)}$ is dominated in the denominator. Let's write out the quantity of interest:

$$\frac{N_*}{\sqrt{n}\Diamond_*} = \frac{\|\boldsymbol{\mu}_{0:k^*}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k^*}} + n\Lambda_*^{-1}\|\boldsymbol{\mu}_{k^*:\infty}\|^2}{\sqrt{\|\boldsymbol{\mu}_{0:k^*}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k^*}} + n^2\Lambda_*^{-2}\|\boldsymbol{\mu}_{k^*:\infty}\|^2_{\boldsymbol{\Sigma}_{k^*:\infty}}}}.$$

Increasing regularization does two things: it changes the value of Λ_*, which serves as a scaling factor in front of the contribution of the tail, and it decreases k^*, therefore recovering the geometry in fewer components. We are going to look at those effects separately.

First, consider the case when k^* doesn't change from changing λ. Note that if the term $\|\boldsymbol{\mu}_{0:k^*}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k^*}}$ dominates in both the numerator and the denominator, then the ratio becomes just $\|\boldsymbol{\mu}_{0:k^*}\|_{\boldsymbol{\Sigma}^{-1}_{0:k^*}}$ up to a constant factor, that is, it is not sensitive to the changes in Λ_*. The same happens if the term $\|\boldsymbol{\mu}_{0:k^*}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k^*}}$ is dominated in both the numerator and the denominator: the ratio becomes $\|\boldsymbol{\mu}_{k^*:\infty}\|^2/\|\boldsymbol{\mu}_{k^*:\infty}\|_{\boldsymbol{\Sigma}_{k^*:\infty}}$, and again it is not sensitive to the changes in Λ_*. Moreover, since we always assume $\Lambda_* > n\lambda_{k^*+1}$ we have

$$n\Lambda_*^{-1}\|\boldsymbol{\mu}_{k^*:\infty}\|^2 \geq n\Lambda_*^{-1}\lambda_{k^*+1}^{-1}\|\boldsymbol{\mu}_{k^*:\infty}\|^2_{\boldsymbol{\Sigma}_{k^*:\infty}} \geq n^2\Lambda_*^{-2}\|\boldsymbol{\mu}_{k^*:\infty}\|^2_{\boldsymbol{\Sigma}_{k^*:\infty}}.$$

Thus, the case in which changing λ can change the bound by more than a constant in this regime is

$$n\Lambda_*^{-1}\|\boldsymbol{\mu}_{k^*:\infty}\|^2 \geq \|\boldsymbol{\mu}_{0:k^*}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k^*}} \geq n^2\Lambda_*^{-2}\|\boldsymbol{\mu}_{k^*:\infty}\|^2_{\boldsymbol{\Sigma}_{k^*:\infty}}.$$

In this case the bound becomes equal to $n\Lambda_*^{-1}\|\boldsymbol{\mu}_{k^*:\infty}\|^2/\|\boldsymbol{\mu}_{0:k^*}\|_{\boldsymbol{\Sigma}_{0:k^*}^{-1}}$ up to a constant factor, so the dependence on Λ_* is inversely proportional. Recall, however, that in order to not change k^* we need to always have $\Lambda_* \le n\lambda_{k^*}$. Putting it together with $\Lambda_* \ge n\lambda_{k^*+1}$ we see that changing regularization in this regime can change the bound by at most $\lambda_{k^*}/\lambda_{k^*+1}$. Thus, there should be a big relative gap between λ_{k^*} and λ_{k^*+1} for that quantity to be large.

The discussion above reveals a recipe for constructing regimes in which increasing regularization can significantly impair the classification accuracy. The formal statement is as follows, its proof can be found in Appendix B.9.

Lemma 53. *For any $\sigma_x \ge 1, L > 1$ there exist constants a, c that only depend on σ_x and absolute constants ε, δ such that the following holds. Suppose that $n > c$, $0 < k < n/c$. Take any $C > 1$ and construct the classification problem as follows:*

1. *Take $\boldsymbol{Z}_{k:\infty}$ with σ_x-sub-Gaussian rows and the sequence $\{\lambda_i\}_{i>k}$ and regularization parameter λ such that $\mathbb{P}(\mathscr{A}_k(L,\lambda)) \ge 1-\delta$ and*
$$\Lambda(\lambda) \ge c\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$
2. *Take $\boldsymbol{Z}_{0:k}$ with σ_x-sub-Gaussian rows independent from $\boldsymbol{Z}_{k:\infty}$, and $\{\lambda_i\}_{i=1}^k$ such that $n\lambda_k \ge C\Lambda(\lambda)$.*
3. *Take $\boldsymbol{\mu}_{k:\infty}$ whose most energy is spread among the eigendirections of $\boldsymbol{\Sigma}$ with small eigenvalues, that is,*
$$\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \le C^{-1}n^{-1}\Lambda(\lambda)\|\boldsymbol{\mu}_{k:\infty}\|^2.$$
4. *Take[4] $\boldsymbol{\mu}_{0:k}$ which balances $\boldsymbol{\mu}_{k:\infty}$ in the following sense:*
$$nC^{-1}\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2 \ge \|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2 \ge n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2. \tag{3.23}$$
5. *Scale $\boldsymbol{\mu}$ up[5] if needed, so it holds that*
$$n\lozenge^2(\lambda) \ge V(\lambda) \quad \text{and } N(\lambda) \ge a\lozenge(\lambda).$$

Then for any λ' such that $\Lambda(\lambda') \ge C\Lambda(\lambda)$
$$\alpha_\varepsilon(\lambda) \ge \frac{C}{c}\alpha_\varepsilon(\lambda').$$

The following corollary, whose proof can be found in Appendix B.9, shows a particular example when the optimal regularization is negative:

[4] Note that such $\boldsymbol{\mu}_{0:k}$ exists because of how we chose $\boldsymbol{\mu}_{k:\infty}$.

[5] Note that the previous conditions were homogeneous in $\boldsymbol{\mu}$, so multiplying it by a scalar does not break them.

Corollary 54. *There exists absolute constants a, b such that the following holds. Take $p = \infty$, $n > a$ and $1 \le k < n/a$. Consider the following classification problem with Gaussian data (in infinite dimension) and no label-flipping noise ($\eta = 0$):*

$$\lambda_i = \begin{cases} 2b, & i \le k, \\ e^{-(i-k)/(bn)}, & i > k. \end{cases}, \quad \mu_i = \begin{cases} 4\sqrt{b/k}, & i \le k, \\ 4\sqrt{b} \cdot 2^{-(i-k)/2}, & i > k. \end{cases}$$

Then the value of λ that maximizes $\alpha_\varepsilon(\lambda)$ is negative.

Increasing regularization can harm by destroying "recovery of the geometry"

Now let's consider a scenario where k^* changes all the way to zero because of increase in regularization. That is, we stop "recovering the geometry" of the first k^* components because of it. For simplicity, consider the case with no tail, that is, $\|\boldsymbol{\mu}_{k^*:\infty}\| = 0$. Informally, increasing regularization will change the classifier from $(\boldsymbol{w}_{\text{ridge}}(\lambda))_{0:k^*} \propto \boldsymbol{\Sigma}_{0:k^*}^{-1}\boldsymbol{\mu}_{0:k^*}$ to $(\boldsymbol{w}_{\text{ridge}}(\lambda'))_{0:k^*} \propto \boldsymbol{\mu}_{0:k^*}$ and the value of $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ will go from $\|\boldsymbol{\mu}_{0:k^*}\|_{\boldsymbol{\Sigma}_{0:k^*}^{-1}}$ to $\|\boldsymbol{\mu}_{0:k^*}\|^2/\|\boldsymbol{\mu}_{0:k^*}\|_{\boldsymbol{\Sigma}_{0:k^*}}$, which may be much smaller depending on $\boldsymbol{\mu}_{0:k}$. This results in the following lemma, whose proof can be found in Appendix B.9:

Lemma 55. *For any $\sigma_x > 1, L > 1$ there exist constants a, c that only depend on L, σ_x and absolute constants ε, δ such that the following holds. Suppose that $n > c$, $0 < k < n/c$. Take any $C > 1$ and construct the classification problem as follows:*

1. *Take $\boldsymbol{Z}_{k:\infty}$ with σ_x-sub-Gaussian rows and the sequence $\{\lambda_i\}_{i>k}$ and regularization parameter λ such that $\mathbb{P}(\mathscr{A}_k(L, \lambda)) \ge 1 - \delta$ and*

$$\Lambda(\lambda) > c\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

2. *Take $\boldsymbol{Z}_{0:k}$ with σ_x-sub-Gaussian rows independent from $\boldsymbol{Z}_{k:\infty}$, and $\{\lambda_i\}_{i=1}^k$ such that $n\lambda_k \ge \Lambda(\lambda)$.*
3. *Take $\boldsymbol{\mu}$ that is only supported on the first k coordinates (i.e., $\|\boldsymbol{\mu}_{k:\infty}\| = 0$) such that*

$$\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}} \ge C\|\boldsymbol{\mu}_{0:k}\|^2. \tag{3.24}$$

4. *Scale $\boldsymbol{\mu}$ up if needed, so that*

$$n\Diamond^2(\lambda) \ge V(\lambda) \quad \text{and } N(\lambda) \ge a\Diamond(\lambda).$$

Then for any λ' such that $\Lambda(\lambda') \geq n\lambda_1$

$$\alpha_\varepsilon(\lambda) \geq \frac{C}{c}\alpha_\varepsilon(\lambda').$$

A natural question is when one can choose such $\boldsymbol{\mu}_{0:k}$ that Equation (3.24) is satisfied. The answer is given by the following.

Lemma 56. *For any $\boldsymbol{\mu}_{0:k} \neq \mathbf{0}_k$*

$$1 \leq \frac{\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}}{\|\boldsymbol{\mu}_{0:k}\|^2} \leq \frac{\lambda_1 + \lambda_k}{2\sqrt{\lambda_1\lambda_k}}.$$

The upper bound is achieved for $\boldsymbol{\mu}_{0:k} = \boldsymbol{e}_1 + \boldsymbol{e}_k$.

Proof. Without loss of generality we can put $\|\boldsymbol{\mu}_{0:k}\|^2 = 1$. Now the numbers $\{\mu_i^2\}$ act as weights: $\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2$ is the weighted average of $\{\lambda_i^{-1}\}_{i=1}^k$ with weights $\{\mu_i^2\}$, while $\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}^{-2}$ is inverse of the weighted average of $\{\lambda_i\}_{i=1}^k$. Thus, for the convex function $f(x) = 1/x$ we can write

$$\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}^2\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2 = \frac{\sum_{i=1}^k \mu_i^2 f(\lambda_i)}{f\left(\sum_{i=1}^k \mu_i^2\lambda_i\right)}.$$

Thus, the lower bound follows from Jensen's inequality. Moreover, if f is a non-negative convex function and X is a random variable with a support $[a, b]$, then the ratio $\mathbb{E}[f(X)]/f(\mathbb{E}[X])$ is maximized by a distribution of X is supported on $\{a, b\}$. That is, we should have $\mu_k^2 = 1 - \mu_1^2$ and $\mu_i = 0$ for $i \notin \{1, k\}$. Now we only need to find the scalar μ_1^2 that maximizes the following:

$$\frac{\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}^2\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2}{\|\boldsymbol{\mu}_{0:k}\|^4} = \left(\lambda_k + (\lambda_1 - \lambda_k)\mu_1^2\right)\left(\lambda_k^{-1} + (\lambda_1^{-1} - \lambda_k^{-1})\mu_1^2\right).$$

Putting the derivative equal to zero yields:

$$0 = (\lambda_1 - \lambda_k)\left(\lambda_k^{-1} + (\lambda_1^{-1} - \lambda_k^{-1})\mu_1^2\right) + \left(\lambda_k + (\lambda_1 - \lambda_k)\mu_1^2\right)(\lambda_1^{-1} - \lambda_k^{-1}),$$
$$\mu_1^2 = 0.5.$$

The maximum value is equal to $\frac{(\lambda_1+\lambda_k)^2}{4\lambda_1\lambda_k}$. □

The following corollary, whose proof can be found in Appendix B.9, shows another example when the optimal regularization is negative:

Corollary 57. *There exist absolute constants $b > c$ such that the following holds. Take $p > bn$, and $b \leq k < n/b$. Consider the following classification problem with Gaussian data (in dimension p) and no label-flipping noise ($\eta = 0$):*

$$\lambda_i = \begin{cases} k^{-4i/k}, & i \leq k, \\ \frac{cn}{pk^4}, & i > k. \end{cases}, \quad \mu_i = \begin{cases} \frac{b\ln(k)}{k^5}\left(\frac{k}{n} + \frac{n}{p}\right), & i \leq k, \\ 0, & i > k. \end{cases}$$

Then the value of λ that maximizes $\alpha_\varepsilon(\lambda)$ is negative.

Regularization with label-flipping noise

Since we don't have a matching upper bound for the case with label-flipping noise, we can only consider the effect of the regularization on the lower bound given in Theorem 42. That bound, up to a constant factor, is given by the following formula:

$$\frac{N - ct\lozenge}{\left([1 + N\sigma_\eta]\sqrt{V + t^2\Delta V} + \lozenge\sqrt{n}\right)}.$$

Let's look at it in the simple regime when t is a constant and $\boldsymbol{\mu}$ is large enough so that $ct\lozenge$ is dominated by N. Thus, we are going to consider the formula

$$\frac{N}{\left([1 + N\sigma_\eta]\sqrt{V} + \lozenge\sqrt{n}\right)}.$$

We can rewrite it up to a constant as a minimum of 2 terms:

$$\frac{1}{\sigma_\eta\sqrt{V}} \wedge \frac{N}{\sqrt{V} + \lozenge\sqrt{n}},$$

and the second term is just the bound for the case $\eta = 0$. We've already seen this in Section 3.4, where we stated that the bound for the case with label-flipping noise goes over the same regimes, and only picks up a new regime for large $\boldsymbol{\mu}$. We already know that increasing λ "doesn't help" in the noiseless regime. It does, however, increase the first term ($V(\lambda)$ is obviously a decreasing function of λ). Thus, regularization can only provide a significant benefit in that new "large $\boldsymbol{\mu}$" regime. Since we don't have a proof of tightness for this bound, however, we leave a more careful study of this effect to future work.

3.7 Related work

Despite the fact that the literature on linear classification in high dimensions is vast, only a few papers studied cases with general covariance structure and the impact of that structure on the prediction accuracy. Here we only review those works, while referring the reader to [34] and [57] for a more broad review of related literature.

Existing results can be split into asymptotic and non-asymptotic. We start with the asymptotic literature. The most common asymptotic regime is the "proportional asymptotic regime," that is, both p and n go to infinity, while their ratio goes to a constant. The results obtained in this regime always require some assumptions on the spectral decay of covariances of the clusters, for example, that all eigenvalues of covariances are bounded from above and below by fixed constants.

[40] consider i.i.d. data $\boldsymbol{x}_i$ from a centered Gaussian distribution with covariance $\boldsymbol{\Sigma}$ in the proportional asymptotic regime. There are two classes, and the probability that a point

$\boldsymbol{x}$ belongs to the first class is a function of $\boldsymbol{x}^\top \boldsymbol{\theta}_*$ for some true parameter vector $\boldsymbol{\theta}_*$. They express the asymptotic classification error of the maximum margin classifier through a solution to a certain system of non-linear equations. [34] consider a problem of multi-class classification of Gaussian mixtures with generic covariances and means in the proportional asymptotic regime. They express asymptotic in-sample and out-of-sample classification errors of a generic convex-loss-minimization algorithm through a solution to a system of non-linear equations under the condition that that solution exists and is unique.

Asymptotic methods often stem from analytical methods from statistical physics, which, even though being mathematically non-rigorous, can accurately predict properties of certain large stochastic systems. [27] consider a binary linear classification problem in which classes have arbitrary means and covariances, and use a non-rigorous computation based on replica-symmetry trick to obtain expressions for the distribution of the solution to a generic loss minimization problem. Interestingly, for the case of symmetric clusters with the same covariance, the empirical mean of that distribution is $(\alpha \boldsymbol{I}_p + \beta \boldsymbol{\Sigma})^{-1}\boldsymbol{\mu}$, where scalars α, β are a solution to a certain system of non-linear equations. Because of this form of the result, one can say that they observed what we call "recovering the geometry". However, the number of components in which the geometry is recovered is hidden behind the system of non-linear equations.

[51] considers a data structure that is very similar to ours. The main result of that paper can be formulated as follows: in an asymptotic setting in which the rows of $\boldsymbol{X}_{k:\infty}$ become very close to orthonormal, the maximum margin solution to the classification problem effectively minimizes hinge loss on the first k components. This paper was an important motivation for our work as it suggested that the regime considered in the regression literature may also lead to fruitful results in classification.

The results of the papers mentioned above are substantially different from ours. All of those papers either consider the maximum margin solution instead of the ridge regression solution, or obtain the result in the form of a solution to a system of equations that is difficult to approach analytically. Because of that, we do not provide more detailed comparisons between our results and the results of those papers. The papers we discuss in the remainder of this section, however, turn out to be directly comparable to ours.

To finish with asymptotic literature, [42] consider a model similar to that of [40], but with a different choice of the covariance structure. Their main result is given for a certain "bi-level" covariance, whose eigenvalues can only take one of two values: there is a small number of eigenvalues with a large value, and a large number of eigenvalues of small value. One can immediately see the similarity between that structure and the structure we introduced in Section 3.1. Even though the goal of [42] was to study the maximum margin solution, the approach they took was motivated by a recent observation that under certain assumptions maximum margin solution coincides with the minimum norm interpolating solution [26, 2]. This phenomenon is known as "support proliferation". Because of that, the main result of [42] is actually derived for the minimum norm interpolating solution, which makes it possible to compare it to our result.

When it comes to non-asymptotic literature, motivated by the same support proliferation idea, [8] studied the classification error minimum norm interpolating solution and [58] studied the ridge regression solution. Finally, [10] obtained a bound on the misclassification probability of maximum margin solution in binary classification. To approach maximum margin solution they utilized its characterization for separable data as the limit of gradient descent on logistic loss. As it turns out, however, assumptions that they consider imply that support proliferation must happen on the event when their proof works.

We will give detailed comparisons with [8, 58, 10, 42]. Before we do that, however, it is worth talking about some similarities that all those papers possess. All four of them had studying the maximum margin solution as their aim. In our notation, the maximum margin solution (MM) is defined as

$$\boldsymbol{w}_{\mathrm{MM}} = \mathrm{argmin}_{\boldsymbol{w}\in\mathbb{R}^p} \|\boldsymbol{w}\| \text{ s.t. } \boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{X}\boldsymbol{w} \geq \mathbf{1}_n, \tag{3.25}$$

where $\boldsymbol{D}_{\hat{\boldsymbol{y}}} = \mathrm{diag}(\hat{\boldsymbol{y}})$. A standard argument with Lagrange multipliers shows that the solution to the optimization problem (3.25) is a conic combination of the columns of the matrix $\boldsymbol{X}^\top\boldsymbol{D}_{\hat{\boldsymbol{y}}}$, and strictly positive coefficients in that conic combination correspond to the inequalities on the right hand side of Equation 3.25 that are saturated, i.e., they are satisfied with equality. The data points (columns of $\boldsymbol{X}^\top$) which correspond to those strictly positive coefficients are called support points. One of the core ideas of [42, 8, 58] is that in some cases "support proliferation" happens with high probability, which means that all points are support points. In this case all inequalities in the constraints become equalities, i.e., $\boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{X}\boldsymbol{w} = \mathbf{1}_n$, and MM coincides with MNI. Motivated by this observation, those papers actually study MNI under support proliferation or in a certain vicinity of that regime. Because of that, our results can be directly compared to the results of those papers.

When it comes to [10], they don't explicitly rely on support proliferation to study MM, but, as we explain in Section 3.7, their proof implies that support proliferation must happen, and thus we can compare our results to theirs too.

Interestingly, one of the conditions under which support proliferation happens is that the whole data distribution has high effective rank. Because of that, most of the results from the above mentioned papers correspond to our results with $k = 0$.

The remainder of this section has the following structure. First, we show that our results generalize the results of [8] and [58]. Then we discuss the relation between our results and those of [10], and show that their bound is weaker then ours for the case of Gaussian data. Finally, we explain how the model considered in [42] is related to ours, and show that some of the conclusions of that paper can be recovered from our analysis too, even though our results do not strictly generalize theirs.

Comparison with "Risk Bounds for Over-parameterized Maximum Margin Classification on Sub-Gaussian Mixtures"

[8] study the same data generating model as ours. Their main result, reformulated in our notation, is given by the following theorem.

Theorem 58 (Theorem 3.1 and Proposition 4.1 from [8])**.** *Suppose the elements of matrix* $\boldsymbol{Z}$ *are independent and* σ_x*-sub-Gaussian, and* $\eta = 0$*. There are constants* C, C' *that only depend on* σ_x *such that the following holds. Assume that*

$$tr(\boldsymbol{\Sigma}) \geq C \max\left\{n^{3/2}\|\boldsymbol{\Sigma}\|, n\|\boldsymbol{\Sigma}\|_F, n\sqrt{\log(n)}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}\right\}, \tag{3.26}$$

and $\|\boldsymbol{\mu}\|^2 \geq C\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$*. Then, with probability at least* $1 - n^{-1}$*,* $\boldsymbol{w}_{MM} = \boldsymbol{w}_{MNI}$ *and*

$$\frac{(\boldsymbol{\mu}^\top \boldsymbol{w}_{MNI})^2}{\|\boldsymbol{w}_{MNI}\|_{\boldsymbol{\Sigma}}^2} \geq C' \frac{n\|\boldsymbol{\mu}\|^4}{n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 + \|\boldsymbol{\Sigma}\|_F^2 + n\|\boldsymbol{\Sigma}\|^2}. \tag{3.27}$$

Note that if we take $k = 0$ and $\lambda = 0$, the assumption imposed in Equation (3.26) implies that $\Lambda \geq n^{3/2}\lambda_1 \gg n$. Moreover, since the data is assumed to be sub-Gaussian, the events $\mathscr{A}_k(L)$ and $\mathscr{B}_k(c_B)$ hold with high probability for constants L, c_B that only depend on σ_x, due to Lemmas 37 and 39. Therefore, our Theorem 42 is applicable and gives the following bound with probability $1 - ce^{-t^2/2}$ (up to a constant factor):

$$\frac{N - ct\lozenge}{\sqrt{V + t^2\Delta V} + \sqrt{n}\lozenge}.$$

Thus, the following proposition, whose proof can be found in Appendix B.10, shows that our bound is at least as good as the bound from [8].

Proposition 59. *Take* $k = 0$ *and some* $c > 1$*. Suppose that* $n\lambda_1 < \Lambda$ *and* $\|\boldsymbol{\mu}\|^2 \geq 2c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$*. Then for* $t < \sqrt{n}$*,*

$$\frac{N - ct\lozenge}{\sqrt{V + t^2\Delta V} + \sqrt{n}\lozenge} \geq \frac{1}{4}\frac{n\|\boldsymbol{\mu}\|^2}{n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} + \sqrt{n}\|\boldsymbol{\Sigma}\|_F + n\|\boldsymbol{\Sigma}\|}. \tag{3.28}$$

Note that the resulting bound does not depend on λ. We have already observed that in Corollary 52: indeed, since $k = 0$, $\boldsymbol{\mu}$ is supported on the tail of the covariance, and regularization does not change the bound by more than a constant factor. Moreover, since they effectively considered $k = 0$, [8] did not observe the effect of "recovering the geometry."

Comparison with "Binary Classification of Gaussian Mixtures: Abundance of Support Vectors, Benign Overfitting and Regularization"

The next paper we compare ours with is [58]. They consider Gaussian $\boldsymbol{Q}$. When it comes to $\boldsymbol{\Sigma}$, they consider two ensembles, which they call "balanced" (see their Definition 2.1) and "bi-level" (see their Definition 2.2). Translating to our terminology, for a balanced ensemble, $k^* = 0$, and for bi-level, $k^* = 1$.

Their result for the balanced ensemble is as follows.

Theorem 60 (Theorem 3 from [58])**.** *There are large absolute constants a, b, c such that the following holds. Assume that rows of $\boldsymbol{Q}$ come from a Gaussian distribution, $k = 0$ and*

$$n\lambda_1 < b \sum_i \lambda_i. \tag{3.29}$$

Take $\lambda \geq 0$. Assume that $\|\boldsymbol{\mu}\|^2 \geq a\,(n\Lambda^{-1}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}})$. Then with probability at least $1 - e^{-n^2/c}$

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} \geq c^{-1} \frac{\|\boldsymbol{\mu}\|^2 - a\,(n\Lambda^{-1}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}})}{(1 \vee n\Lambda^{-1}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}})\sqrt{\sum_i \lambda_i^2} + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}. \tag{3.30}$$

We see that their bound is at most within a constant factor of

$$\frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\Sigma}\|_F + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \frac{n\|\boldsymbol{\Sigma}\|_F}{\lambda + \mathrm{tr}(\boldsymbol{\Sigma})}}.$$

Comparing to our bound from Equation 3.28, we see that the bound from [58] has $\|\boldsymbol{\Sigma}\|_F$ in the denominator, which is larger (up to a constant) than $\|\boldsymbol{\Sigma}\|_F/\sqrt{n} + \|\boldsymbol{\Sigma}\|$ that stands in the denominator of Equation 3.28 (after dividing both numerator and denominator by n). Moreover, it picks up an additional term $\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \frac{n\|\boldsymbol{\Sigma}\|_F}{\lambda + \mathrm{tr}(\boldsymbol{\Sigma})}$ in the denominator. Thus, the bound from Theorem 60 is worse than the one from Equation 3.28. Note that, just as in the previous section, Equation (3.29) implies that our Theorem 42 is applicable with high probability. That is, Proposition 59 shows that our result generalizes the result for balanced ensembles from [58].

When it comes to the bi-level ensemble, the result of [58] translated to our notation is given by the following theorem.

Theorem 61 (Theorem 5 from [58])**.** *There are large absolute constants a, b, c such that the following holds. Assume that rows of $\boldsymbol{Q}$ come from a Gaussian distribution. Take $\boldsymbol{\mu}$ that is supported on one coordinate, i.e., $\boldsymbol{\mu} = \mu_j \boldsymbol{e}_j$ for some j, and $j > 1$. Assume that $\|\boldsymbol{\mu}\|^2 \geq a\,(n\Lambda^{-1}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}})$. Assume that $k = 1$, $\lambda \geq 0$ and*

$$bn\lambda_1 > \sum_{i>1} \lambda_i \quad \text{and} \quad bn\lambda_2 < \sum_{i>2} \lambda_i. \tag{3.31}$$

Denote

$$A = \lambda_1 \frac{\Lambda + n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}{n\lambda_1 + \Lambda}, \quad B = (1 + n\Lambda^{-1}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}})\sqrt{\sum_{i \neq 1, j} \lambda_i^2}.$$

Then with probability at least $1 - e^{-n/c}$,

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} \geq c^{-1} \frac{\|\boldsymbol{\mu}\|^2(1 - cn\Lambda^{-1}\lambda_j) - c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}{A + B + \lambda_j + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}. \tag{3.32}$$

Note that for $k = 1$ the second part of Equation (3.31) yields $\Lambda > b\lambda_{k+1}$. Thus, if b is large enough, under that assumption, the events $\mathscr{A}_k(L)$ and $\mathscr{B}_k(c_B)$ hold with high probability for absolute constants L, c_B, due to Lemmas 37 and 39. Therefore, our Theorem 42 is applicable just as it was in Section 3.7. So, the following proposition, whose proof can be found in Appendix B.10, shows that our bound generalizes the bound for bi-level ensembles from [58].

Proposition 62. *Take $k = 1$ and some $c > 1$. Assume that $\lambda > 0$, $n\lambda_{k+1} \le \sum_{i>k}\lambda_i$, $\|\boldsymbol{\mu}_{0:k}\| = 0$, and $\|\boldsymbol{\mu}\|^2 \ge 2c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$. Take any $j > 1$ and define A, B as in Theorem 61. Then for $t \le \sqrt{n}$*

$$\frac{N - ct\lozenge}{\sqrt{V + t^2\Delta V} + \sqrt{n}\lozenge} \ge \frac{1}{6}\frac{\|\boldsymbol{\mu}\|^2}{A + B + \lambda_j + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}.$$

Overall, we see that the bounds from [58] are not sharp. Moreover, since they considered either $k = 0$ or $k = 1$ and $\boldsymbol{\mu}$ supported on a single coordinate, they did not observe the effect of "recovering the geometry."

Comparison with "Finite-sample analysis of interpolating linear classifiers in the overparameterized regime"

[10] consider almost the same data generating model as ours: two clusters with symmetric means and the same covariances. Only their definition of the noise is different: they consider arbitrary corruptions of the distribution of $(\boldsymbol{x}, y)$ that preserve the marginal distribution of $\boldsymbol{x}$ and have bounded total variation distance with the initial distribution. Label-flipping noise can be seen as a particular case of such corruption.

Nevertheless, comparing our results with those of [10] is not straightforward for two reasons. First, they consider the MM solution, while we consider the ridge and MNI solutions. Second, [10] impose assumptions that are incomparable with ours, for example they assume that elements of $\boldsymbol{Q}$ have bounded sub-Gaussian norms, while we only have proofs that the events $\mathscr{A}_k(L)$ and $\mathscr{B}_k(c_B)$ hold with high probability when elements of $\boldsymbol{Z}$ are sub-Gaussian (see our Lemmas 37 and 39).

Regarding the first potential issue, we note that they in fact consider a regime where the max-margin solution coincides with MNI. To see this, note that by Lemma A.2 of [16], for max-margin to coincide with MNI, it suffices for the training data to be 'nearly-orthogonal' in the sense that $\|\boldsymbol{x}_k\|^2 \gg n \max_{i,j} \frac{\|\boldsymbol{x}_i\|^2}{\|\boldsymbol{x}_j\|^2} \max_{i\ne j} |\boldsymbol{x}_i^\top \boldsymbol{x}_j|$ for every training sample $(\boldsymbol{x}_k, y_k)$. One can verify this property holds in their setting by using their Lemma 10 together with their assumption (A.3).

To alleviate the problem with the differences in the assumptions, we compare the results for the case of Gaussian distributions, where both our and their results are directly applicable. We also only consider the label-flipping noise here, since it is a particular case of the noise considered in [10]. When translated into our notation, that result is given by the following.

Theorem 63 (Theorem 1 from [10], Gaussian case)**.** *Fix some constant $\kappa \in (0,1)$. Suppose rows of $\boldsymbol{Q}$ are i.i.d. samples from a Gaussian distribution. Suppose that $\lambda_i \leq 1$ for every $i \in \{1,\dots,p\}$ and $\sum_i \lambda_i \geq \kappa p$. There is a constant c that only depends on κ and an absolute constant b such that the following holds.*

Take $\delta \in (e^{-n/c}, c^{-1})$. Assume that $p \geq cn^2\log(n/\delta)$, $p/(cn) \geq \|\boldsymbol{\mu}\|^2 \geq c\log(n/\delta)$, and $\eta \leq 1/c$. Then with probability $1-\delta$ over the draw of $\boldsymbol{X}, \hat{\boldsymbol{y}}$

$$\mathbb{P}_{(\boldsymbol{x},\hat{y})}(\hat{y}\boldsymbol{x}^\top \boldsymbol{w}_{MM} < 0) \leq \eta + \exp\left(-b\frac{\|\boldsymbol{\mu}\|^4}{p}\right), \tag{3.33}$$

where $(\boldsymbol{x}, \hat{y})$ is a new data point from the data distribution with label-flipping noise.

Note that assumptions of Theorem 63 yield for $n > e/c$:

$$\sum_i \lambda_i \geq \kappa p \geq cn^2\log(nc) \geq cn\lambda_1.$$

Thus, for $k = 0$ we have $\Lambda > cn\lambda_{k+1}$. According to Lemmas 37 and 39, if c is a large enough absolute constant, both events $\mathscr{A}_k(L)$ and $\mathscr{B}_k(c_B)$ hold with probability at least $1 - ce^{-n/c}$ for some absolute constants L and c_B. Thus, Theorem 42 is applicable for $k = 0$, and yields the result with probability $1 - ce^{-n/c} - ce^{-t^2/2}$. To match the probability of $1-\delta$ from Theorem 63 we should take $t = \sqrt{2\log(1/\delta)}$. Finally, in Section 3.1 we saw that in the Gaussian case the error probability on a new noiseless point $(\boldsymbol{x}, y)$ is $\Phi(-\boldsymbol{\mu}^\top\boldsymbol{w}/\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}})$. Since $\Phi(-z) \leq e^{-z^2/2}$ for every $z > 0$, to recover the result of Theorem 63 we just need to show that $\boldsymbol{\mu}^\top\boldsymbol{w}/\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}} \gtrsim \|\boldsymbol{\mu}\|^2/\sqrt{p}$. Thus, the following proposition, whose proof can be found in Appendix B.10, shows that our result is stronger than Theorem 63.

Proposition 64. *Assume that $\lambda_i \leq 1$ for any i and $\sum_{i=1}^p \lambda_i \geq \kappa p$ for some constant $\kappa \in (0,1]$. Take $k = 0$, $\lambda = 0$ and some $c > 1$. Suppose additionally that $\kappa p/n \geq \|\boldsymbol{\mu}\|^2 \geq (2ct)^2/(\kappa^2 n)$, and $t^2 < n\kappa$.*

Then

$$\frac{N - ct\Diamond}{[1 + N\sigma_\eta]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n}} \geq \frac{1}{10}\frac{\|\boldsymbol{\mu}\|^2\sqrt{n\kappa}}{\sqrt{p}}.$$

That is, our lower bound picks up an additional factor of $\sqrt{n}$ compared to the bound from Theorem 63.

Comparison with "Classification vs regression in overparameterized regimes: Does the loss function matter?"

Apart from the data generating model considered in this chapter, there is another model that was recently considered in the literature on linear classification. Consider a centered Gaussian distribution with covariance $\boldsymbol{\Sigma}$. When a point $\boldsymbol{\xi}$ is generated from this distribution, it gets assigned the label $\operatorname{sign}(\boldsymbol{\xi}^\top\boldsymbol{\alpha})$ for some vector $\boldsymbol{\alpha}\in\mathbb{R}^p$. Thus, the domain is split into two clusters. It is easy to see that the centers of the clusters are

$$\mathbb{E}[\operatorname{sign}(\boldsymbol{\xi}^\top\boldsymbol{\alpha})\boldsymbol{\xi}] = \mathbb{E}[\operatorname{sign}(\boldsymbol{z}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha})\boldsymbol{\Sigma}^{1/2}\boldsymbol{z}] = \boldsymbol{\Sigma}^{1/2}\cdot\sqrt{\frac{2}{\pi}}\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}\|} = \sqrt{\frac{2}{\pi\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{\alpha}}}\boldsymbol{\Sigma}\boldsymbol{\alpha} =: \boldsymbol{m},$$

where we used $\boldsymbol{z}$ to denote a vector from the isotropic Gaussian distribution, and denoted the centers of the clusters as $\pm\boldsymbol{m}$, which plays the role of $\boldsymbol{\mu}$. The covariance within a cluster is not $\boldsymbol{\Sigma}$, but a rank-one correction to it, namely

$$\begin{aligned}\boldsymbol{\Sigma}' = \mathbb{E}\left[(\boldsymbol{\xi}\operatorname{sign}(\boldsymbol{\xi}^\top\boldsymbol{\alpha}) - \boldsymbol{m})(\operatorname{sign}(\boldsymbol{\xi}^\top\boldsymbol{\alpha})\boldsymbol{\xi})^\top\right] = \mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] - \boldsymbol{m}\mathbb{E}[\operatorname{sign}(\boldsymbol{\xi}^\top\boldsymbol{\alpha})\boldsymbol{\xi}]^\top = \\ = \boldsymbol{\Sigma} - \boldsymbol{m}\boldsymbol{m}^\top = \boldsymbol{\Sigma} - \frac{2\boldsymbol{\Sigma}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}}{\pi\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{\alpha}} = \boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{I}_p - \frac{2}{\pi}\boldsymbol{P}_{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}}\right)\boldsymbol{\Sigma}^{1/2},\end{aligned}$$

where we denoted the projection on the direction of $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}$ as $\boldsymbol{P}_{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}}$. Because of the factor $2/\pi < 1$ in front of it, the matrix $\boldsymbol{I}_p - \frac{2}{\pi}\boldsymbol{P}_{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}}$ is still within a constant factor of identity, so the covariance of a cluster is within a constant factor of $\boldsymbol{\Sigma}$.

Moreover, for a classifier $\boldsymbol{\xi}\to\operatorname{sign}(\boldsymbol{\xi}^\top\boldsymbol{w})$, the probability to assign a wrong label is

$$\mathbb{P}\left(\operatorname{sign}(\boldsymbol{z}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}) \neq \operatorname{sign}(\boldsymbol{z}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha})\right) = \frac{\angle(\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha})}{\pi} = \frac{1}{\pi}\arccos\left(\frac{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{w}}{\|\boldsymbol{\alpha}\|_{\boldsymbol{\Sigma}}\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}}\right),$$

where we used $\angle(\cdot,\cdot)$ to denote the angle between two vectors. Note that the argument of arccos is almost the same as the quantity $\boldsymbol{\mu}^\top\boldsymbol{w}/\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}$ studied in this chapter: indeed, plugging in the formulas for the mean and the covariance of the cluster we obtain

$$\frac{\boldsymbol{m}^\top\boldsymbol{w}}{\sqrt{\boldsymbol{w}\boldsymbol{\Sigma}'\boldsymbol{w}}} = \sqrt{\frac{2}{\pi}}\frac{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{w}}{\sqrt{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{\alpha}}\sqrt{\boldsymbol{w}\boldsymbol{\Sigma}'\boldsymbol{w}}},$$

and we saw that $(1-2/\pi)\boldsymbol{w}\boldsymbol{\Sigma}\boldsymbol{w} \le \boldsymbol{w}\boldsymbol{\Sigma}'\boldsymbol{w} \le \boldsymbol{w}\boldsymbol{\Sigma}\boldsymbol{w}$.

Thus, in principle, our results can apply directly to this model. The caveat, however, is that our bounds are only defined up to a constant multiplier, while the quantity $\frac{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{w}}{\|\boldsymbol{\alpha}\|_{\boldsymbol{\Sigma}}\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}}$ is always between minus one and one. For example, our bounds cannot distinguish between perfect classification and some constant probability of error that is less than 0.5.

Now let's compare our results with the result of [42], who consider such a model. The main result of [42] is their Theorem 13, which considers the following construction: there are three non-negative real-valued parameters q, r, s such that $r < 1 < s$, $q < s - r$. The covariance is diagonal, that is, $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$, and $p = n^s$. The spectrum of $\boldsymbol{\Sigma}$ has a bi-level structure, that is

$$\lambda_i = \begin{cases} n^{s-q-r}, & \text{for } i \le n^r, \\ (1-n^{-q})/(1-n^{r-s}) & \text{for } i > n^r. \end{cases} \tag{3.34}$$

Finally, [42] consider $\boldsymbol{\alpha} = \boldsymbol{e}_1$ (note that, similarly to [58], taking such $\boldsymbol{\alpha}$ hides the effect of "recovering the geometry," since $\boldsymbol{\alpha}$ has the same direction as $\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}$). For this choice of $\boldsymbol{\alpha}$, the mean of the positive cluster becomes $\boldsymbol{m} = \sqrt{\frac{2}{\pi\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{\alpha}}}\boldsymbol{\Sigma}\boldsymbol{\alpha} = \sqrt{\frac{2\lambda_1}{\pi}}\boldsymbol{e}_1$.

[42] consider the asymptotic setting with n approaching infinity, and compute the classification error of the MNI. Namely, their Theorem 13 shows that if $q+r < (s+1)/2$ then the misclassification probability approaches zero, while for $q + r > (s + 1)/2$ it approaches 0.5. That is, the quantity $\frac{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{w}}{\|\boldsymbol{\alpha}\|_{\boldsymbol{\Sigma}}\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}}$ approaches 1 when $q+r < (s+1)/2$, and 0 if $q+r > (s+1)/2$.

Proposition 65 below shows that one can see the same phase transition in our results. However, as our bounds are only defined up to a constant multiplier, we do not recover the result of [42] precisely.

Proposition 65. *Take real* q, r, s *such that* $0 \le r < 1 < s$, $0 \le q < s - r$. *Consider* $p = n^s$, $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$, *and* $\boldsymbol{\mu} = \sqrt{2\lambda_1/\pi}\boldsymbol{e}_1$, *where* $\{\lambda_i\}_{i=1}^p$ *are given by Equation* (3.34). *Take* $\lambda = 0$, $k = n^r$, *and* c *to be any constant that doesn't depend on* n.

Then, as n *goes to infinity, for* $t < n^{0.499r}$ *the following holds:*

$$\frac{N - ct\Diamond}{\sqrt{V + t^2\Delta V} + \sqrt{n}\Diamond} = (1 + o_n(1))\frac{N}{\sqrt{V} + \sqrt{n}\Diamond} = \begin{cases} o_n(1), & 2q + 2r - 1 - s > 0, \\ \frac{1+o_n(1)}{\sqrt{2\pi}} & 2q + 2r - 1 - s = 0, \\ \sqrt{\frac{2}{\pi}} + o_n(1) & 2q + 2r - 1 - s < 0. \end{cases}$$

Here we use $o_n(1)$ *to denote quantities that converge to zero as* n *goes to infinity.*

3.8 Conclusions and further directions

In this chapter we studied classification accuracy of the ridge regression solution in a binary classification problem. We derived tight bounds for the case without label-flipping noise, and a lower bound for the case with label-flipping noise. Our bounds are additionally supported by geometric derivations for the minimum norm interpolating solution, which explain the structure of the solution vector. Even though we don't provide a matching upper bound for the case with label-flipping noise, the geometric derivations show that the vector $\boldsymbol{Q}\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ plays an important role, thus suggesting that the term $\sigma_\eta\sqrt{V}$ should indeed appear in the bound for $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, and that our bound is indeed tight.

Our bounds yield several novel qualitative conclusions. We discover the effect of "recovering the geometry" in the first k^* components, which was seemingly missed in the previous literature. For the setting without label-flipping noise, we show that there is no benefit (in a certain sense) of increasing regularization beyond the point where the (regularized) covariance obtains a tail of high effective rank, and that the optimal regularization can even be negative. When it comes to the case with label flipping noise and benign overfitting, we discover that the bound for this case exhibits the same behavior as the bound for the noiseless case, unless $\boldsymbol{\mu}$ is large in magnitude. In the latter case, our bound loses dependence on $\boldsymbol{\mu}$ completely, and the conditions for benign overfitting in this regime coincide with the conditions from the regression setting considered in Chapter 2.

Despite all the above mentioned progress, there are still gaps in our understanding of benign overfitting in this model, which we leave for future work. The most obvious task is to obtain the matching upper bound for the case with label-flipping noise. As explained above, we believe that our bound should be tight, at least when η is a constant. The dependence of the bound on η, however, is probably not sharp when η becomes small. This is because our argument relies on sub-Gaussianity of a Bernoulli random variable with parameter η, but when that parameter is small, the Bernoulli random variable behaves as a heavy-tailed one. Thus, the argument using sub-Gaussianity may not be sharp.

Next, our argument only works if there exists k for which the tail of the covariance has high effective rank. However, the bound that we obtained suggests that this structure may be necessary for benign overfitting to occur. Indeed, as we explained above, the sufficient conditions for benign condition that we obtain are very similar to those for regression, and we showed in Section 2.3 that high effective rank in the tail of the covariance is necessary for benign overfitting in regression. Proving the necessity of this regime in classification is another direction of future work.

Finally, even though we use a very similar regime for both regression and classification, and there are a lot of technical similarities between the results, we do not have a high-level explanation of benign overfitting that would unify the regression and classification settings. Resolving this, and understanding when and how the noise that is interpolated in training does not impact classification accuracy, are important directions for future work.

Bibliography

[1] Pedro Abdalla and Nikita Zhivotovskiy. "Covariance Estimation: Optimal Dimension-free Guarantees for Adversarial Corruption and Heavy Tails". In: *arXiv:2205.08494* (2023).

[2] Navid Ardeshir, Clayton Sanford, and Daniel J Hsu. "Support vector machines and linear regression coincide with very high-dimensional features". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 4907–4918. URL: `https://proceedings.neurips.cc/paper_files/paper/2021/file/26d4b4313a7e5828856bc0791fca39a2-Paper.pdf`.

[3] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* (2020). ISSN: 0027-8424. DOI: `10.1073/pnas.1907378117`. eprint: `https://www.pnas.org/content/early/2020/04/22/1907378117.full.pdf`. URL: `https://www.pnas.org/content/early/2020/04/22/1907378117`.

[4] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. "Deep learning: a statistical viewpoint". In: *ArXiv* abs/2103.09177 (2021).

[5] Mikhail Belkin, Daniel Hsu, and Ji Xu. "Two models of double descent for weak features". In: *ArXiv* abs/1903.07571 (2019).

[6] Koby Bibas, Yaniv Fogel, and Meir Feder. "A New Look at an Old Problem: A Universal Learning Approach to Linear Regression". In: July 2019, pp. 2304–2308. DOI: `10.1109/ISIT.2019.8849398`.

[7] Florentina Bunea, Seth Strimas-Mackey, and Marten H Wegkamp. "Interpolating Predictors in High-Dimensional Factor Regression." In: *J. Mach. Learn. Res.* 23 (2022), pp. 10–1.

[8] Yuan Cao, Quanquan Gu, and Mikhail Belkin. "Risk Bounds for Over-parameterized Maximum Margin Classification on Sub-Gaussian Mixtures". In: *CoRR* abs/2104.13628 (2021). arXiv: `2104.13628`. URL: `https://arxiv.org/abs/2104.13628`.

[9] Michael Celentano, Theodor Misiakiewicz, and Andrea Montanari. *Minimum complexity interpolation in random features models*. 2021. arXiv: `2103.15996 [cs.LG]`.

[10] Niladri S. Chatterji and Philip M. Long. “Finite-sample analysis of interpolating linear classifiers in the overparameterized regime”. In: *Journal of Machine Learning Research* 22.129 (2021), pp. 1–30.

[11] Chen Cheng and Andrea Montanari. *Dimension free ridge regression.* 2022. arXiv: `2210.08571 [math.ST]`.

[12] Geoffrey Chinot and Matthieu Lerasle. “On the robustness of the minimum ℓ_2 interpolator”. In: *ArXiv* abs/2003.05838 (2021).

[13] Michał Dereziński, Feynman Liang, Zhenyu Liao, and Michael W. Mahoney. “Precise expressions for random projections: Low-rank approximation and randomized Newton”. In: *ArXiv* abs/2006.10653 (2020).

[14] Michał Dereziński, Feynman Liang, and Michael Mahoney. “Exact expressions for double descent and implicit regularization via surrogate random design”. Dec. 2019.

[15] Edgar Dobriban and Stefan Wager. “High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification”. In: *arXiv: Statistics Theory* (2015), pp. 247–279.

[16] Spencer Frei, Gal Vardi, Peter L. Bartlett, and Nathan Srebro. “Benign Overfitting in Linear Classifiers and Leaky ReLU Networks from KKT Conditions for Margin Maximization”. In: *Conference on Learning Theory.* 2023.

[17] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. “Linearized two-layers neural networks in high dimension”. In: *ArXiv* abs/1904.12191 (2020).

[18] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. “When Do Neural Networks Outperform Kernel Methods?” In: *ArXiv* abs/2006.13409 (2020).

[19] Nikhil Ghosh, Song Mei, and Bin Yu. *The Three Stages of Learning Dynamics in High-Dimensional Kernel Methods.* 2021. arXiv: `2111.07167 [stat.ML]`.

[20] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* Third. The Johns Hopkins University Press, 1996.

[21] Olivier Guédon, Alexander Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. “On the interval of fluctuation of the singular values of random matrices”. In: *Journal of the European Mathematical Society* 19.5 (2017), pp. 1469–1505. DOI: `10.4171/jems/697`.

[22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. “Surprises in high-dimensional ridgeless least squares interpolation”. In: *Annals of statistics* 50.2 (2022), p. 949.

[23] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[24] Trevor J. Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. “Surprises in High-Dimensional Ridgeless Least Squares Interpolation”. In: *ArXiv* (2019). URL: `https://arxiv.org/abs/1903.08560v4`.

[25] Daniel Hsu, Sham M. Kakade, and Tong Zhang. “Random design analysis of ridge regression”. In: *Foundations of Computational Mathematics* 14.3 (2014), pp. 569–600.

[26] Daniel Hsu, Vidya Muthukumar, and Ji Xu. “On the proliferation of support vectors in high dimensions”. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2021, pp. 91–99.

[27] Hanwen Huang and Qinglong Yang. “Large dimensional analysis of general margin based classification methods”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.11 (Nov. 2021), p. 113401. DOI: `10.1088/1742-5468/ac2edd`. URL: `https://dx.doi.org/10.1088/1742-5468/ac2edd`.

[28] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. “Optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization.” In: *arXiv: Statistics Theory* (2020).

[29] Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. “Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds and Benign Overfitting”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. 2021. URL: `https://openreview.net/forum?id=FyOhThdDBM`.

[30] Vladimir Koltchinskii and Karim Lounici. “Concentration inequalities and moment bounds for sample covariance operators”. In: *Bernoulli* 23.1 (2017), pp. 110–133. DOI: `10.3150/15-BEJ730`. URL: `https://doi.org/10.3150/15-BEJ730`.

[31] Tengyuan Liang and Alexander Rakhlin. “Just Interpolate: Kernel ”Ridgeless” Regression Can Generalize”. In: *ArXiv* abs/1808.00387 (2018).

[32] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. “On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels”. In: *ArXiv* abs/1908.10292 (2020).

[33] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. “On the Risk of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels”. In: *ArXiv* abs/1908.10292 (2019).

[34] Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. “Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 10144–10157.

[35] Andrew D. McRae, Santhosh Karnik, Mark A. Davenport, and Vidya Muthukumar. *Harmless interpolation in regression and classification with structured features*. 2021. arXiv: `2111.05198 [stat.ML]`.

[36] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. “Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration”. In: *ArXiv* abs/2101.10588 (2021).

[37] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. *Learning with invariances in random features and kernel models.* 2021. arXiv: `2102.13219 [stat.ML]`.

[38] Song Mei and Andrea Montanari. “The generalization error of random features regression: Precise asymptotics and the double descent curve”. In: *Communications on Pure and Applied Mathematics* (2019).

[39] Theodor Misiakiewicz and Song Mei. *Learning with convolution and pooling operations in kernel methods.* 2021. arXiv: `2111.08308 [stat.ML]`.

[40] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. “The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime”. In: *Preprint, arXiv:1911.01544* (2023).

[41] Andrea Montanari and Yiqiao Zhong. “The interpolation phase transition in neural networks: Memorization and generalization under lazy training”. In: *The Annals of Statistics* 50.5 (2022), pp. 2816–2847. DOI: `10.1214/22-AOS2211`. URL: `https://doi.org/10.1214/22-AOS2211`.

[42] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. “Classification vs regression in overparameterized regimes: Does the loss function matter?” In: *Journal of Machine Learning Research* 22.222 (2021), pp. 1–69.

[43] Vidya Muthukumar, Kailas Vodrahalli, and Anant Sahai. “Harmless interpolation of noisy data in regression”. In: *2019 IEEE International Symposium on Information Theory (ISIT)* (2019), pp. 2299–2303.

[44] Preetum Nakkiran. “More Data Can Hurt for Linear Regression: Sample-wise Double Descent”. In: *ArXiv* abs/1912.07242 (2019).

[45] Adhyyan Narang, Vidya Muthukumar, and Anant Sahai. *Classification and Adversarial examples in an Overparameterized Linear Model: A Signal Processing Perspective.* 2021. arXiv: `2109.13215 [cs.LG]`.

[46] Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. “In Defense of Uniform Convergence: Generalization via Derandomization with an Application to Interpolating Predictors”. In: *Proceedings of the 37th International Conference on Machine Learning.* Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 7263–7272. URL: `http://proceedings.mlr.press/v119/negrea20a.html`.

[47] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. “Asymptotics of Ridge (less) Regression under General Source Condition”. In: *ArXiv* abs/2006.06386 (2020).

[48] Mark Rudelson and Roman Vershynin. “Hanson-Wright inequality and sub-gaussian concentration”. In: *Electronic Communications in Probability* 18.none (2013), pp. 1–9. DOI: `10.1214/ECP.v18-2865`. URL: `https://doi.org/10.1214/ECP.v18-2865`.

[49] Mark Rudelson and Roman Vershynin. "Sampling from large matrices: An approach through geometric functional analysis". In: *J. ACM* 54.4 (July 2007), 21–es. ISSN: 0004-5411. DOI: 10.1145/1255443.1255449. URL: https://doi.org/10.1145/1255443.1255449.

[50] Ohad Shamir. *The Implicit Bias of Benign Overfitting*. 2022. arXiv: 2201.11489 [cs.LG].

[51] Ohad Shamir. "The Implicit Bias of Benign Overfitting". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 448–478. URL: https://proceedings.mlr.press/v178/shamir22a.html.

[52] Konstantin Tikhomirov. "Sample Covariance Matrices of Heavy-Tailed Distributions". In: *International Mathematics Research Notices* 2018.20 (Apr. 2017), pp. 6254–6289. ISSN: 1073-7928. DOI: 10.1093/imrn/rnx067. eprint: https://academic.oup.com/imrn/article-pdf/2018/20/6254/26127510/rnx067.pdf. URL: https://doi.org/10.1093/imrn/rnx067.

[53] Alexander Tsigler and Peter L. Bartlett. "Benign overfitting in ridge regression". In: *ArXiv* abs/2009.14286 (2020).

[54] Alexander Tsigler and Peter L. Bartlett. "Benign overfitting in ridge regression". In: *Journal of Machine Learning Research* 24.123 (2023), pp. 1–76. URL: http://jmlr.org/papers/v24/22-1398.html.

[55] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. DOI: 10.1017/9781108231596.

[56] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed Sensing: Theory and Applications*. Ed. by Yonina C. Eldar and GittaEditors Kutyniok. Cambridge University Press, 2012, pp. 210–268. DOI: 10.1017/CBO9780511794308.006.

[57] Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. "Benign Overfitting in Multiclass Classification: All Roads Lead to Interpolation". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. 2021. URL: https://openreview.net/forum?id=OO5jpovbdHO.

[58] Ke Wang and Christos Thrampoulidis. "Binary Classification of Gaussian Mixtures: Abundance of Support Vectors, Benign Overfitting and Regularization". In: *Preprint, arXiv:2011.09148* (2021).

[59] Denny Wu and Ji Xu. "On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression". In: *ArXiv* abs/2006.05800 (2020).

[60] Ji Xu and Daniel J Hsu. "On the number of variables to use in principal component regression". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 5094–5103. URL: `http://papers.nips.cc/paper/8753-on-the-number-of-variables-to-use-in-principal-component-regression.pdf`.

[61] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations (ICLR)*. 2017.

[62] Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. "On Uniform Convergence and Low-Norm Interpolation Learning". In: *ArXiv* abs/2006.05942 (2021).

Appendix A

Proofs for Chapter 2

A.1 Ridge regression

We are interested in evaluating the MSE of the ridge estimator. For positive regularization parameter λ that estimator is defined as

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \hat{\boldsymbol{\theta}}(\boldsymbol{y}) =& \operatorname{argmin}_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \right\} \\ =& \left(\lambda \boldsymbol{I}_p + \boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}.\end{aligned}$$

In the overparametrized case (i.e., $p > n$), however, the latter expression has a singularity at zero, because the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ does not have full rank. If $\lambda = 0$ the solution to the minimization problem above is not unique. Moreover, if $\lambda < 0$, no solution exists at all because we are minimizing a quadratic form whose matrix has negative singular values. To alleviate these issues and extend the definition of the solution to non-positive values of λ, we propose the following: since the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ doesn't have full rank, we can apply the Sherman-Morrison-Woodbury formula:

$$\left(\lambda \boldsymbol{I}_p + \boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \lambda^{-1} \boldsymbol{I}_p - \lambda^{-2} \boldsymbol{X}^\top (\boldsymbol{I}_n + \lambda^{-1} \boldsymbol{X}\boldsymbol{X}^\top)^{-1} \boldsymbol{X}.$$

So,

$$\begin{aligned}\left(\lambda \boldsymbol{I}_p + \boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top =& \lambda^{-1} \boldsymbol{X}^\top - \lambda^{-2} \boldsymbol{X}^\top (\boldsymbol{I}_n + \lambda^{-1} \boldsymbol{X}\boldsymbol{X}^\top)^{-1} \boldsymbol{X}\boldsymbol{X}^\top \\ =& \lambda^{-1} \boldsymbol{X}^\top - \lambda^{-1} \boldsymbol{X}^\top (\boldsymbol{I}_n + \lambda^{-1} \boldsymbol{X}\boldsymbol{X}^\top)^{-1} (\lambda^{-1} \boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{I}_n - \boldsymbol{I}_n) \\ =& \lambda^{-1} \boldsymbol{X}^\top (\boldsymbol{I}_n + \lambda^{-1} \boldsymbol{X}\boldsymbol{X}^\top)^{-1}, \\ \hat{\boldsymbol{\theta}}(\boldsymbol{y}) =& \boldsymbol{X}^\top (\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1} \boldsymbol{y}.\end{aligned}$$

The matrix $\boldsymbol{X}\boldsymbol{X}^\top$ has full rank, and the expression above is continuous in λ as long as $\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_n$ stays PD. When $\lambda = 0$, $\boldsymbol{X}^\top (\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1} y$ is the minimum norm interpolating solution. Therefore, we use the expression

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y}) := \boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$$

to define the ridge regression solution for any $\lambda > -\mu_n(\boldsymbol{X}\boldsymbol{X}^\top)$.

Note that $\hat{\boldsymbol{\theta}}(y)$ is linear in y. Since we have $y = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ we can also write

$$\hat{\boldsymbol{\theta}}(y) = \hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) + \hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}).$$

The first term is the noiseless estimate; its error gives the bias term. The second term is the estimate obtained when the signal is pure noise. It gives the variance term.

For the full MSE we have

$$\begin{aligned}\|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 =& \|\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) + \hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2\\ \le& 2\|\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 + 2\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2\\ =& 2(B + V_{\boldsymbol{\varepsilon}}),\end{aligned}$$

where we introduced bias B and variance $V_{\boldsymbol{\varepsilon}}$:

$$\begin{aligned}B :=& \|\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 = \|(\boldsymbol{I}_p - \boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X})\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2,\\ V_{\boldsymbol{\varepsilon}} :=& \|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2 \qquad\qquad = \|\boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\varepsilon}\|_{\boldsymbol{\Sigma}}^2.\end{aligned}$$

Finally, since $V_{\boldsymbol{\varepsilon}}$ is a quadratic form in $\boldsymbol{\varepsilon}$, by Lemma 82 if the noise is sub-Gaussian, then its value is controlled by its expectation with high probability. That expectation, in its turn, scales linearly with the variance $v_{\boldsymbol{\varepsilon}}^2$ of the noise. Therefore, we can decouple the effect of the noise and only study the following purified variance term:

$$\begin{aligned}V :=& \frac{1}{v_{\boldsymbol{\varepsilon}}^2}\mathbb{E}_{\boldsymbol{\varepsilon}} V_{\boldsymbol{\varepsilon}}\\ =& \operatorname{tr}((\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1})\\ =& \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-2}\boldsymbol{X}).\end{aligned}$$

The main aim of our work is to give sharp non-asymptotic bounds for B and V.

A.2 Proofs for the ridgeless derivation

Algebraic decompositions

Lemma 66. *Suppose $k < n$, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is an invertible matrix, and $\boldsymbol{V} \in \mathbb{R}^{n\times k}$ is such that $\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{A}$ is invertible. Then*

$$\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{A})^{-2}\boldsymbol{V} = (\boldsymbol{I}_k + \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-2}\boldsymbol{V}(\boldsymbol{I}_k + \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{V})^{-1}.$$

Proof. We use the Sherman–Morrison–Woodbury formula to write

$$(\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{A})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{V}(\boldsymbol{I}_k + \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-1}. \tag{A.1}$$

Denote $\boldsymbol{M}_1 := \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{V}$ and $\boldsymbol{M}_2 := \boldsymbol{V}^\top\boldsymbol{A}^{-2}\boldsymbol{V}$. Applying (A.1), we get

$$\begin{aligned}
&\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{A})^{-2}\boldsymbol{V}\\
&= \boldsymbol{V}^\top\Big(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{V}(\boldsymbol{I}_k + \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-1}\Big)^2\boldsymbol{V}\\
&= \boldsymbol{V}^\top\Big(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{V}(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-1}\Big)^2\boldsymbol{V}\\
&= \boldsymbol{V}^\top\Big(\boldsymbol{A}^{-2} - \boldsymbol{A}^{-2}\boldsymbol{V}(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{V}(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-2}\\
&\qquad + \boldsymbol{A}^{-1}\boldsymbol{V}(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-2}\boldsymbol{V}(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-1}\Big)\boldsymbol{V}\\
&= \boldsymbol{M}_2 - \boldsymbol{M}_2(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_1 - \boldsymbol{M}_1(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_2\\
&\qquad + \boldsymbol{M}_1(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_2(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_1\\
&= \boldsymbol{M}_2 - \boldsymbol{M}_2(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_1 - \boldsymbol{M}_1(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_2(\boldsymbol{I}_k - (\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_1)\\
&= \boldsymbol{M}_2(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1} - \boldsymbol{M}_1(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_2(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\\
&= (\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_2(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1},
\end{aligned}$$

where we used the identity $\boldsymbol{I}_k - (\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}\boldsymbol{M}_1 = (\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}$ twice in the second last equality and the identity $\boldsymbol{I}_k - \boldsymbol{M}_1(\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1} = (\boldsymbol{I}_k + \boldsymbol{M}_1)^{-1}$ in the last equality. □

Lemma 3. *For any $i \in \{1, \dots, p\}$ define $\boldsymbol{A}_{-i} := \sum_{j \neq i} \lambda_j \boldsymbol{z}_j\boldsymbol{z}_j^\top$. If $\boldsymbol{A}_{-i}$ is invertible, then*

$$\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}^{-2}\boldsymbol{z}_i = \frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2}\boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2}.$$

Proof. We use Lemma 66, which is a consequence of the Sherman-Woodbury-Morrison formula.

$$\begin{aligned}
\lambda_i^2 \boldsymbol{z}_i^\top \left(\sum_j \lambda_j \boldsymbol{z}_j\boldsymbol{z}_j^\top\right)^{-2} \boldsymbol{z}\boldsymbol{z}_i &= \lambda_i^2 \boldsymbol{z}_i^\top \left(\lambda_i \boldsymbol{z}_i\boldsymbol{z}_i^\top + \boldsymbol{A}_{-i}\right)^{-2}\boldsymbol{z}_i\\
&= \frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2}\boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2},
\end{aligned}$$

by Lemma 66, for the case $k = 1$ and $\boldsymbol{V} = \sqrt{\lambda_i}\boldsymbol{z}_i$. □

Concentration inequalities

We use some standard results about sub-Gaussian and sub-Exponential random variables. First of all, we define sub-Exponentiality:

Definition 67 (Definition 2.7.5 from [55]). *For any centered random variable v we define its sub-Exponential norm as*

$$\|v\|_{\psi_1} := \inf\left\{t > 0 : \mathbb{E}\exp(|v|/t) \le 2\right\}.$$

If $\|v\|_{\psi_1} \le \sigma$, we say that the distribution of v is σ-sub-Exponential.

We are going to need the following direct consequence of Propositions 2.5.2 and 2.7.1 and Lemma 2.7.6 from [55]:

Lemma 68. *There is a universal constant c such that for any random variable ξ that is centered, σ-sub-Gaussian, and unit variance, ξ^2-1 is a centered $c\sigma^2$-sub-Exponential random variable.*

Second, we are going to use the following form of Bernstein's inequality, which is Theorem 2.8.2 in [55]:

Lemma 69. *There is a universal constant c such that, for any independent, mean zero, σ-sub-Exponential random variables $\xi_1, \dots, \xi_N$, any $a = (a_1, \dots, a_N) \in \mathbb{R}^n$, and any $t \ge 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i\xi_i\right| > t\right) \le 2\exp\left[-c\min\left(\frac{t^2}{\sigma^2\sum_{i=1}^N a_i^2}, \frac{t}{\sigma\max_{1\le i\le n} a_i}\right)\right].$$

Corollary 70. *There is a universal constant c such that for any non-increasing sequence $\{\lambda_i\}_{i=1}^\infty$ of non-negative numbers such that $\sum_{i=1}^\infty \lambda_i < \infty$, and any independent, centered, σ-sub-Exponential random variables $\{\xi_i\}_{i=1}^\infty$, and any $x > 0$, with probability at least $1-2e^{-x}$*

$$\left|\sum_i \lambda_i\xi_i\right| \le c\sigma\max\left(x\lambda_1, \sqrt{x\sum_i\lambda_i^2}\right).$$

Proof. Denote the constant from Lemma 69 as c_1. Plug in the following value of t in the result of that lemma:

$$t = \sigma\max\left(c_1^{-1}x\max_i a_i, \sqrt{c_1^{-1}x\sum_{i=1}^N a_i^2}\right).$$

Finally, change notation from $(a_i)_{i=1}^N$ to $(\lambda_i)_{i=1}^p$ and take $c = \max(c_1^{-1}, c_1^{-1/2})$. □

Concentration of a quadratic form evaluated at a vector with independent components is implied by the following lemma.

Lemma 71 (A version of Hanson-Wright inequality)**.** *Suppose $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is a (possibly random) matrix, and $\boldsymbol{z} \in \mathbb{R}^n$ is a random vector with independent components, that have unit variances and are σ-sub-Gaussian. If $\boldsymbol{z}$ is independent from $\boldsymbol{M}$, then for an absolute constant c and any $t > 0$ with probability $1 - 2e^{-t}$,*

$$|\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z} - tr(\boldsymbol{M})| \le c\sigma^2(\sqrt{t}\|\boldsymbol{M}\|_F + t\|\boldsymbol{M}\|).$$

Proof. By Theorem 6.2.1 (Hanson-Wright inequality) in [55], for some absolute constant c_1 for any $t > 0$,

$$\mathbb{P}\left\{|\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z} - \mathbb{E}\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z}| \ge t\right\} \le 2\exp\left(-c_1 \min\left\{\frac{t^2}{\|\boldsymbol{M}\|_F^2\sigma^4}, \frac{t}{\|\boldsymbol{M}\|\sigma^2}\right\}\right).$$

Substituting t by $\sigma^2 \max(\sqrt{t/c_1}\|\boldsymbol{M}\|_F, t\|\boldsymbol{M}\|/c_1)$ yields the result. □

Corollary 72. *There is an absolute constant c such that for any centered random vector $\boldsymbol{z} \in \mathbb{R}^n$ with independent σ-sub-Gaussian coordinates with unit variances, any random subspace $\mathscr{L}$ of $\mathbb{R}^n$ of codimension k that is independent of $\boldsymbol{z}$, and any $t > 0$, with probability at least $1 - 4e^{-t}$,*

$$\begin{aligned}\|\boldsymbol{z}\|^2 \le& n + c\sigma^2(t + \sqrt{nt}),\\ \|\Pi_{\mathscr{L}}\boldsymbol{z}\|^2 \ge& n - c\sigma^2(k + t + \sqrt{nt}),\end{aligned}$$

where $\Pi_{\mathscr{L}}$ is the orthogonal projection on $\mathscr{L}$.

Proof. Denote the coordinates of $\boldsymbol{z}$ as $(z_i)_{i=1}^n$, that is, $\boldsymbol{z} = (z_1, \dots, z_n)^\top$. First of all, since $\|\boldsymbol{z}\|^2 = \sum_{i=1}^n z_i^2$ — a sum of n σ^2-sub-Exponential random variables, by Corollary 70, for some absolute constant c and for any $t > 0$, with probability at least $1 - 2e^{-t}$,

$$\left|\|\boldsymbol{z}\|^2 - n\right| \le c\sigma^2 \max(t, \sqrt{nt}).$$

Second, we can write

$$\|\Pi_{\mathscr{L}}\boldsymbol{z}\|^2 = \|\boldsymbol{z}\|^2 - \|\Pi_{\mathscr{L}^\perp}\boldsymbol{z}\|^2.$$

Recall that projectorsare self-adjoing operators, so the matrix $\Pi_{\mathscr{L}^\perp}$ is symmetric PSD. Since $\|\Pi_{\mathscr{L}^\perp}\| = 1$ and $\operatorname{tr}(\Pi_{\mathscr{L}^\perp}) = \operatorname{tr}(\Pi_{\mathscr{L}^\perp}^2) = k$, by Lemma 71, for some absolute constant c_1 with probability at least $1 - 2e^{-t}$,

$$\begin{aligned}\|\Pi_{\mathscr{L}^\perp}\boldsymbol{z}\|^2 =& \boldsymbol{z}^\top \Pi_{\mathscr{L}^\perp}\boldsymbol{z}\\ \le& k + c_1\sigma^2(t + \sqrt{kt})\\ \le& c_1\sigma^2(2k + 2t).\end{aligned}$$

Thus, with probability at least $1 - 4e^{-t}$

$$\begin{aligned}\|\boldsymbol{z}\|^2 \le& n + c\sigma^2 \max(t, \sqrt{nt}),\\ \|\Pi_{\mathscr{L}}\boldsymbol{z}\|^2 \ge& \|\boldsymbol{z}\| - 2c_1\sigma^2(k + t)\\ \ge& n - c_2\sigma^2(k + t + \max(t, \sqrt{nt})),\end{aligned}$$

where we chose a new large enough absolute constant c_2 in the last transition. □

Epsilon-net argument

Lemma 73 (ϵ-net argument)**.** *Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a symmetric matrix, and $\mathcal{N}_\epsilon$ is an ϵ-net on the unit sphere $\mathcal{S}^{n-1}$ in the Euclidean norm, where $\epsilon < \frac{1}{2}$. Then*

$$\|\boldsymbol{A}\| \leq (1-\epsilon)^{-2} \max_{\boldsymbol{u}\in\mathcal{N}_\epsilon} |\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}|.$$

Proof. Denote the eigenvalues of $\boldsymbol{A}$ as $\lambda_1, \ldots, \lambda_n$ and assume $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Denote the first eigenvector of $\boldsymbol{A}$ as $\boldsymbol{v} \in \mathcal{S}^{n-1}$, and take $\Delta\boldsymbol{v} \in \mathbb{R}^n$ such that $\boldsymbol{v} + \Delta\boldsymbol{v} \in \mathcal{N}_\epsilon$ and $\|\Delta\boldsymbol{v}\| \leq \epsilon$. Denote the coordinates of $\Delta\boldsymbol{v}$ in the eigenbasis of $\boldsymbol{A}$ as $\Delta v_1, \ldots, \Delta v_n$. Now we can write

$$\begin{aligned}
\left|(\boldsymbol{v}+\Delta\boldsymbol{v})^\top \boldsymbol{A}(\boldsymbol{v}+\Delta\boldsymbol{v})\right| &= \left|\lambda_1 + 2\lambda_1\Delta v_1 + \sum_{i=1}^n \lambda_i \Delta v_i^2\right| \\
&= |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 + \sum_{i=2}^n \frac{\lambda_i}{\lambda_1}\Delta v_i^2\right| \\
&\geq |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 - \sum_{i=2}^n \Delta v_i^2\right| \\
&= |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 - \|\Delta\boldsymbol{v}\|^2 + \Delta v_1^2\right| \\
&= |\lambda_1| \cdot \left|1 + 2\left(\Delta v_1 + \Delta v_1^2\right) - \|\Delta\boldsymbol{v}\|^2\right| \\
&\geq |\lambda_1| \cdot \left|1 + 2\left(-\|\Delta\boldsymbol{v}\| + (-\|\Delta\boldsymbol{v}\|)^2\right) - \|\Delta\boldsymbol{v}\|^2\right| \\
&= |\lambda_1| \cdot \left|1 - 2\|\Delta\boldsymbol{v}\| + \|\Delta\boldsymbol{v}\|^2\right| \\
&\geq |\lambda_1| \cdot |1 - 2\epsilon + \epsilon^2| \\
&= \|\boldsymbol{A}\|(1-\epsilon)^2,
\end{aligned}$$

where the first inequality holds because the λ_is are decreasing in magnitude, and the last two inequalities hold since the functions $x + x^2$ and $2x + x^2$ are both increasing on $(-\frac{1}{2}, \infty)$ and $\Delta v_1 \geq -\|\Delta v\| \geq -\epsilon \geq -\frac{1}{2}$. □

Lemma 4. *Set $\lambda = 0$. Suppose all elements of matrix $\boldsymbol{Z}$ are independent and σ_x-sub-Gaussian. There is a constant c that only depends on σ_x such that with probability* $1 - 2e^{-n}$

$$\mu_n(\boldsymbol{A}) \geq \sum_{i=1}^p \lambda_i - c\left(n\lambda_1 + \sqrt{n\sum_{i=1}^p \lambda_i^2}\right) = tr(\boldsymbol{\Sigma}) - c(n\|\boldsymbol{\Sigma}\| + \sqrt{n}\|\boldsymbol{\Sigma}\|_F),$$

$$\mu_1(\boldsymbol{A}) \leq \sum_{i=1}^p \lambda_i + c\left(n\lambda_1 + \sqrt{n\sum_{i=1}^p \lambda_i^2}\right) = tr(\boldsymbol{\Sigma}) + c(n\|\boldsymbol{\Sigma}\| + \sqrt{n}\|\boldsymbol{\Sigma}\|_F).$$

Proof. For a fixed vector $\boldsymbol{v} \in \mathbb{R}^n$, Proposition 2.6.1 from [55] implies that for some constant c_1 and any i the random variable $\boldsymbol{v}^\top \boldsymbol{z}_i$ is $c_1\|\boldsymbol{v}\|^2\sigma_x$-sub-Gaussian. Thus, for any fixed unit vector $\boldsymbol{v}$, as $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} = \sum_i \lambda_i(\boldsymbol{v}^\top \boldsymbol{z}_i)^2$, Lemma 68 and Corollary 70 imply that for some constant c_2 with probability at least $1-2e^{-t}$,

$$\left|\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} - \sum \lambda_i\right| \le c_2\sigma_x^2 \max\left(\lambda_1 t, \sqrt{t\sum \lambda_i^2}\right).$$

Let $\mathcal{N}$ be a $\frac{1}{4}$-net on the sphere $\mathcal{S}^{n-1}$ with respect to the Euclidean distance such that $|\mathcal{N}| \le 9^n$. Applying the union bound over the elements of $\mathcal{N}$, we see that with probability $1-2e^{-t}$, every $v \in \mathcal{N}$ satisfies

$$\left|\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} - \sum \lambda_i\right| \le c_2\sigma_x^2 \max\left(\lambda_1(t+n\ln 9), \sqrt{(t+n\ln 9)\sum_i \lambda_i^2}\right).$$

Since $\mathcal{N}$ is a $\frac{1}{4}$-net, by Lemma 73, we need to multiply the quantity above by $(1-1/4)^{-2}$ to get the bound on the norm of the $\boldsymbol{A} - \boldsymbol{I}_n\sum_i \lambda_i$. Denote

$$\Diamond = \left(\lambda_1(t+n\ln 9) + \sqrt{(t+n\ln 9)\sum_i \lambda_i^2}\right).$$

Thus, with probability at least $1-2e^{-t}$,

$$\left\|\boldsymbol{A} - \boldsymbol{I}_n\sum_i \lambda_i\right\| \le c_3\sigma_x^2\Diamond.$$

Taking $t=n$ finishes the proof. □

Eigenvalues of low rank corrections

For symmetric matrices $\boldsymbol{U}, \boldsymbol{V}$ we use the notation $\boldsymbol{U} \preceq \boldsymbol{V}$ to denote that the matrix $\boldsymbol{V}-\boldsymbol{U}$ is PSD.

Recall (half of) the Courant-Fischer-Weyl theorem.

Lemma 74. *For any symmetric $n \times n$ matrix $\boldsymbol{A}$, and any $i \in [n]$, $\mu_i(\boldsymbol{A})$ is the minimum, over all subspaces $\mathscr{U}$ of $\mathbb{R}^n$ of dimension $n-i$, of the maximum, over all unit-length $\boldsymbol{u} \in \mathscr{L}$, of $\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}$.*

Lemma 75 (Monotonicity of eigenvalues)**.** *If symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfy $\boldsymbol{A} \preceq \boldsymbol{B}$, then, for any $i \in [n]$, we have $\mu_i(\boldsymbol{A}) \le \mu_i(\boldsymbol{B})$.*

Proof. Let $\mathscr{U}$ be the subspace of $\mathbb{R}^n$ of dimension $n-i$ that minimizes the maximum over all unit-length $\boldsymbol{u} \in \mathscr{U}$, of $\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u}$, and let $\mathscr{V}$ be the analogous subspace for $\boldsymbol{B}$. We have

$$\begin{aligned}
\mu_i(\boldsymbol{A}) &= \max_{\boldsymbol{u} \in \mathscr{U}:\|\boldsymbol{u}\|=1} \boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u} \quad \text{(by Lemma 74)} \\
&\leq \max_{\boldsymbol{v} \in \mathscr{V}:\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top \boldsymbol{A} \boldsymbol{v} \quad \text{(since } \mathscr{U} \text{ is the minimizer)} \\
&\leq \max_{\boldsymbol{v} \in \mathscr{V}:\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top \boldsymbol{B} \boldsymbol{v} \quad \text{(since } \boldsymbol{A} \preceq \boldsymbol{B}) \\
&= \mu_i(\boldsymbol{B}),
\end{aligned}$$

by Lemma 74, completing the proof. □

Lemma 7. *1. for all* $i \geq 1$,

$$\mu_{k+1}(\boldsymbol{A}_{-i}) \leq \mu_{k+1}(\boldsymbol{A}) \leq \mu_1(\boldsymbol{A}_k),$$

2. for all $1 \leq i \leq k$,

$$\mu_n(\boldsymbol{A}) \geq \mu_n(\boldsymbol{A}_{-i}) \geq \mu_n\left(\boldsymbol{A}_k\right),$$

Proof. First, the matrix $\boldsymbol{A} - \boldsymbol{A}_k$ has rank at most k (as a sum of k matrices of rank 1). Thus, there is a linear space $\mathscr{L}$ of dimension $n-k$ such that for all $\boldsymbol{v} \in \mathscr{L}$, $\boldsymbol{v}^\top \boldsymbol{A} \boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{A}_k \boldsymbol{v} \leq \mu_1(\boldsymbol{A}_k)\|\boldsymbol{v}\|^2$, and so $\mu_{k+1}(\boldsymbol{A}) \leq \mu_1(\boldsymbol{A}_k)$.

Second, by the Courant-Fischer-Weyl Theorem, for all i and j, $\mu_j(\boldsymbol{A}_{-i}) \leq \mu_j(\boldsymbol{A})$ (see Lemma 75). On the other hand, for $i \leq k$, $\boldsymbol{A}_k \preceq \boldsymbol{A}_{-i}$, so all the eigenvalues of $\boldsymbol{A}_{-i}$ are lower bounded by $\mu_n(\boldsymbol{A}_k)$. □

Proof of the upper bound

Lemma 8. *Suppose all elements of matrix* $\boldsymbol{Z}$ *are independent and* σ_x*-sub-Gaussian. There are constants* $b, c \geq 1$ *that only depend on* σ_x *such that if* $0 \leq k \leq n/c$, $r_k \geq bn$, *and* $l \leq k$ *then with probability at least* $1 - 8e^{-n/c}$,

$$V \leq c\left(\frac{l}{n} + \frac{n\sum_{i>l}\lambda_i^2}{(\lambda_{k+1} r_k)^2}\right).$$

Proof. By Lemma 3,

$$\begin{aligned}
V &= \sum_i \lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}^{-2} \boldsymbol{z}_i \\
&= \sum_{i=1}^{l} \frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2} + \sum_{i>l} \lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}^{-2} \boldsymbol{z}_i. \qquad \text{(A.2)}
\end{aligned}$$

First, consider the sum up to l. Take b to be equal to the constant c from Lemma 6. If $r_k \geq bn$, Lemmas 6 and 7 show that with probability at least $1 - 2e^{-n}$, for all $i \leq k$,

$\mu_n(\boldsymbol{A}_{-i}) \geq \lambda_{k+1}r_k/c_1$, and, for all i, $\mu_{k+1}(\boldsymbol{A}_{-i}) \leq c_1\lambda_{k+1}r_k$. The lower bounds on the $\mu_n(\boldsymbol{A}_{-i})$'s imply that, for all $\boldsymbol{z} \in \mathbb{R}^n$ and $1 \leq i \leq l$,

$$\boldsymbol{z}^\top \boldsymbol{A}_{-i}^{-2}\boldsymbol{z} \leq \frac{c_1^2\|\boldsymbol{z}\|^2}{(\lambda_{k+1}r_k)^2},$$

and the upper bounds on the $\mu_{k+1}(\boldsymbol{A}_{-i})$'s give

$$\boldsymbol{z}^\top \boldsymbol{A}_{-i}^{-1}\boldsymbol{z} \geq (\Pi_{\mathscr{L}_i}\boldsymbol{z})^\top \boldsymbol{A}_{-i}^{-1}\Pi_{\mathscr{L}_i}\boldsymbol{z} \geq \frac{\|\Pi_{\mathscr{L}_i}\boldsymbol{z}\|^2}{c_1\lambda_{k+1}r_k},$$

where $\mathscr{L}_i$ is the span of the $n-k$ eigenvectors of $\boldsymbol{A}_{-i}$ corresponding to its smallest $n-k$ eigenvalues. So for $i \leq l$,

$$\frac{\lambda_i^2\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2}\boldsymbol{z}_i}{(1+\lambda_i\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2} \leq \frac{\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2}\boldsymbol{z}_i}{(\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2} \leq c_1^4\frac{\|\boldsymbol{z}_i\|^2}{\|\Pi_{\mathscr{L}_i}\boldsymbol{z}_i\|^4}. \tag{A.3}$$

Next, we apply Corollary 72 l times, together with a union bound, to show that with probability at least $1-4e^{-t}$, for all $1 \leq i \leq l$,

$$\|\boldsymbol{z}_i\|^2 \leq n + a\sigma_x^2(t+\ln k+\sqrt{n(t+\ln k)}) \leq c_2 n, \tag{A.4}$$

$$\|\Pi_{\mathscr{L}_i}\boldsymbol{z}_i\|^2 \geq n - a\sigma_x^2(k+t+\ln k+\sqrt{n(t+\ln k)}) \geq n/c_3, \tag{A.5}$$

provided that $t < n/c_0$ and $c > c_0$ for some sufficiently large c_0 (note that c_2 and c_3 only depend on c_0, a and σ_x, and we can still take c large enough in the end without changing c_2 and c_3). Combining (A.3), (A.4), and (A.5), with probability at least $1-5e^{-n/c_0}$,

$$\sum_{i=1}^{l}\frac{\lambda_i^2\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2}\boldsymbol{z}_i}{(1+\lambda_i\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2} \leq c_4\frac{l}{n}.$$

Second, consider the second sum in (A.2). Lemma 7 shows that, on the same high probability event that we considered in bounding the first half of the sum, $\mu_n(\boldsymbol{A}) \geq \lambda_{k+1}r_k/c_1$. Hence,

$$\sum_{i>l}\lambda_i^2\boldsymbol{z}_i^\top \boldsymbol{A}^{-2}\boldsymbol{z}_i \leq \frac{c_1^2\sum_{i>l}\lambda_i^2\|\boldsymbol{z}_i\|^2}{(\lambda_{k+1}r_k)^2}.$$

Notice that $\sum_{i>l}\lambda_i^2\|\boldsymbol{z}_i\|^2$ is a weighted sum of σ_x^2-sub-Exponential random variables, with the weights given by the λ_i^2 in blocks of size n. Corollary 70 implies that, with probability at least $1-2e^{-t}$,

$$\begin{aligned}\sum_{i>l}\lambda_i^2\|\boldsymbol{z}_i\|^2 &\leq n\sum_{i>l}\lambda_i^2 + a\sigma_x^2\max\left(\lambda_{l+1}^2 t, \sqrt{tn\sum_{i>l}\lambda_i^4}\right)\\ &\leq n\sum_{i>l}\lambda_i^2 + a\sigma_x^2\max\left(t\sum_{i>l}\lambda_i^2, \sqrt{tn}\sum_{i>l}\lambda_i^2\right)\\ &\leq c_5 n\sum_{i>l}\lambda_i^2,\end{aligned}$$

because $t < n/c_0$. Combining the above gives

$$\sum_{i>l} \lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}^{-2} \boldsymbol{z}_i \le c_6 n \frac{\sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k)^2}.$$

Finally, putting both parts together and taking $c > \max\{c_0, c_4, c_6\}$ gives the lemma. □

Proof of the lower bound

Lemma 76. *There is a constant c such that for any $i \ge 1$ with $\lambda_i > 0$, and any $0 \le k \le n/c$, with probability at least $1 - 6e^{-n/c}$,*

$$\frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2} \ge \frac{1}{cn} \left(1 + \frac{\sum_{j>k} \lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2}.$$

Proof. Fix $i \ge 1$ with $\lambda_i > 0$ and $0 \le k \le n/c$. Apply Lemma 4 to matrix $\boldsymbol{A}_k$ instead of $\boldsymbol{A}$, and note that $\sqrt{n}\|\boldsymbol{\Sigma}_{k:\infty}\|_F \le \sqrt{n\|\boldsymbol{\Sigma}_{k:\infty}\|\mathrm{tr}(\boldsymbol{\Sigma}_{k:\infty})\|} \le n\|\boldsymbol{\Sigma}_{k:\infty}\| + \mathrm{tr}(\boldsymbol{\Sigma}_{k:\infty})$. Plugging the resulting bound on $\mu_1(\boldsymbol{A}_k)$ into Lemma 7 shows that with probability at least $1 - 2e^{-n}$,

$$\mu_{k+1}(\boldsymbol{A}_{-i}) \le c_1 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right),$$

and hence

$$\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i \ge \frac{\|\Pi_{\mathscr{L}_i} \boldsymbol{z}_i\|^2}{c_1 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right)}.$$

By Corollary 72, with probability at least $1 - 4e^{-t}$,

$$\|\Pi_{\mathscr{L}_i} \boldsymbol{z}_i\|^2 \ge n - a\sigma_x^2 (k + t + \sqrt{tn}) \ge n/c_2,$$

provided that $t < n/c_0$ and $c > c_0$ for some sufficiently large c_0. Thus, with probability at least $1 - 5e^{-n/c_3}$,

$$\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i \ge \frac{n}{c_3 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right)},$$

hence

$$1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i \le \left(\frac{c_3 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right)}{\lambda_i n} + 1\right) \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i.$$

Dividing $\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i$ by the square of both sides, we have

$$\frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(1+\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2} \geq \left(\frac{c_3 \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n \right)}{\lambda_i n} + 1 \right)^{-2} \frac{\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2}.$$

Also, from the Cauchy-Schwarz inequality and Corollary 72 again, we have that on the same event,

$$\begin{aligned} \frac{\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i)^2} &\geq \frac{\boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{\left\| \boldsymbol{A}_{-i}^{-1} \boldsymbol{z}_i \right\|^2 \|\boldsymbol{z}_i\|^2} \\ &= \frac{1}{\|\boldsymbol{z}_i\|^2} \geq \frac{1}{n + a\sigma_x^2 (t + \sqrt{nt})} \geq \frac{1}{c_4 n}. \end{aligned}$$

Choosing c suitably large gives the lemma.

□

Lemma 77. *Suppose that $\{\eta_i\}_{i=1}^p$ is a sequence of non-negative random variables, and that $\{t_i\}_{i=1}^p$ is a sequence of non-negative real numbers (at least one of which is strictly positive) such that, for some $\delta \in (0,1)$ for any $i \leq p$ with probability at least $1-\delta$, $\eta_i > t_i$. Then with probability at least $1-2\delta$,*

$$\sum_{i=1}^n \eta_i \geq \frac{1}{2} \sum_{i=1}^p t_i.$$

Proof. We know that, for all $i \leq p$, $\mathbb{P}(\eta_i > t_i) \geq 1-\delta$. Consider the following event:

$$\mathscr{E} = \left\{ \sum_{i=1}^p \eta_i < \frac{1}{2} \sum_{i=1}^p t_i \right\},$$

and denote its probability as $\mathbb{P}(\mathscr{E}) = c\delta$ for some $c \in (0, \delta^{-1})$. On the one hand, by the definition of the event, we have

$$\frac{1}{\mathbb{P}(\mathscr{E})} \mathbb{E} \left[1_{\mathscr{E}} \sum_{i=1}^p \eta_i \right] \leq \frac{1}{2} \sum_{i=1}^p t_i,$$

where $1_{\mathscr{E}}$ is the indicator of the event $\mathscr{E}$. On the other hand, note that for any i,

$$\begin{aligned} \mathbb{E}[\eta_i 1_{\mathscr{E}}] \geq & \mathbb{E}[t_i 1_{\{\eta_i \geq t_i\} \cap \mathscr{E}}] \\ &= t_i \mathbb{P}(\{\eta_i \geq t_i\} \cap \mathscr{E}) \\ &\geq t_i (\mathbb{P}\{\eta_i \geq t_i\} + \mathbb{P}(\mathscr{E}) - 1) \\ &\geq t_i (c-1)\delta. \end{aligned}$$

So

$$\mathbb{E}\left[1_{\mathscr{E}}\sum_{i=1}^{p}\eta_i\right] \geq (c-1)\delta\sum_{i=1}^{p}t_i,$$
$$\frac{1}{\mathbb{P}(\mathscr{E})}\mathbb{E}\left[1_{\mathscr{E}}\sum_{i=1}^{p}\eta_i\right] \geq (1-c^{-1})\sum_{i=1}^{p}t_i.$$

Thus, we obtain

$$\frac{1}{2}\sum_{i=1}^{p}t_i \geq (1-c^{-1})\sum_{i=1}^{p}t_i,$$
$$c \leq 2,$$
$$\mathbb{P}\left(\sum_{i=1}^{p}\eta_i < \frac{1}{2}\sum_{i=1}^{p}t_i\right) = \mathbb{P}(\mathscr{E}) = c\delta \leq 2\delta.$$

□

Lemma 9. *Suppose all elements of matrix $\boldsymbol{Z}$ are independent and σ_x-sub-Gaussian. There is a constant c that only depends on σ_x such that for any $0 \leq k \leq n/c$ and any $b > 1$ with probability at least $1 - 10e^{-n/c}$,*

1. *If $r_k < bn$, then $V \geq \frac{k+1}{cb^2 n}$.*
2. *If $r_k \geq bn$, then*
$$V \geq \frac{1}{cb^2}\min_{l\leq k}\left(\frac{l}{n} + \frac{b^2 n\sum_{i>l}\lambda_i^2}{\left(\lambda_{k+1}r_k\right)^2}\right).$$

In particular, if all choices of $k \leq n/c$ give $r_k < bn$, then $r_{n/c} < bn$ implies that with probability at least $1 - 12e^{-n/c}$, $V \geq (cb)^{-2}$—at least a constant.

Proof. From Lemmas 3, 76 and 77, with probability at least $1 - 12e^{-n/c_1}$,

$$\begin{aligned}
V &\geq \frac{1}{c_1 n}\sum_i\left(1 + \frac{\sum_{j>k}\lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2} \\
&\geq \frac{1}{c_2 n}\sum_i \min\left\{1, \frac{n^2\lambda_i^2}{\left(\sum_{j>k}\lambda_j\right)^2}, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\} \\
&\geq \frac{1}{c_2 b^2 n}\sum_i \min\left\{1, \left(\frac{bn}{r_k}\right)^2\frac{\lambda_i^2}{\lambda_{k+1}^2}, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\}.
\end{aligned}$$

Now, if $r_k < bn$, then the second term in the minimum is always bigger than the third term, and in that case,

$$V \geq \frac{1}{c_2 b^2 n} \sum_i \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\} \geq \frac{k+1}{c_2 b^2 n}.$$

On the other hand, if $r_k(\lambda) \geq bn$,

$$\begin{aligned} V &\geq \frac{1}{c_2 b^2} \sum_i \min\left\{\frac{1}{n}, \frac{b^2 n \lambda_i^2}{(\lambda_{k+1} r_k)^2}\right\} \\ &= \frac{1}{c_2 b^2} \min_{l \leq k} \left(\frac{l}{n} + \frac{b^2 n \sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k)^2}\right), \end{aligned}$$

where the equality follows from the fact that the λ_is are non-increasing. □

Choosing ℓ

Lemma 10. *For any $b \geq 1$ and $k^* := \min\{k : r_k \geq bn\}$, if $k^* < \infty$, we have*

$$\min_{l \leq k^*} \left(\frac{l}{bn} + \frac{bn \sum_{i>l} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*})^2}\right) = \frac{k^*}{bn} + \frac{bn \sum_{i>k^*} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*})^2} = \frac{k^*}{bn} + \frac{bn}{R_{k^*}},$$

where we introduced $R_k := \left(\sum_{i>k} \lambda_i\right)^2 / \left(\sum_{i>k} \lambda_i^2\right)$.

Proof. We can write the function of l being minimized as

$$\begin{aligned} \frac{l}{bn} + \frac{bn \sum_{i>l} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*})^2} &= \sum_{i=1}^{l} \frac{1}{bn} + \sum_{i>l} \frac{bn \lambda_i^2}{(\lambda_{k^*+1} r_{k^*})^2} \\ &\geq \sum_{i=1}^{k^*} \min\left\{\frac{1}{bn}, \frac{bn \lambda_i^2}{(\lambda_{k^*+1} r_{k^*})^2}\right\} \\ &\quad + \sum_{i>k^*} \frac{bn \lambda_i^2}{(\lambda_{k^*+1} r_{k^*})^2} \\ &= \sum_{i=1}^{l^*} \frac{1}{bn} + \sum_{i>l^*} \frac{bn \lambda_i^2}{(\lambda_{k^*+1} r_{k^*})^2}, \end{aligned}$$

where l^* is the largest value of $i \leq k^*$ for which

$$\frac{1}{bn} \leq \frac{bn \lambda_i^2}{(\lambda_{k^*+1} r_{k^*})^2},$$

since the λ_i^2 are non-increasing. This condition holds iff

$$\lambda_i \geq \frac{\lambda_{k^*+1} r_{k^*}}{bn}.$$

The definition of k^* implies $r_{k^*-1} < bn$. So we can write

$$\begin{aligned} r_{k^*} &= \frac{\sum_{i>k^*} \lambda_i}{\lambda_{k^*+1}} \\ &= \frac{\sum_{i>k^*-1} \lambda_i - \lambda_{k^*}}{\lambda_{k^*+1}} \\ &= \frac{\lambda_{k^*}}{\lambda_{k^*+1}} (r_{k^*-1} - 1) \\ &< \frac{\lambda_{k^*}}{\lambda_{k^*+1}} (bn - 1), \end{aligned}$$

and so the minimizing l is k^*. Also,

$$\frac{\sum_{i>k^*} \lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}\right)^2} = \frac{\sum_{i>k^*} \lambda_i^2}{\left(\sum_{i>k^*} \lambda_i\right)^2} = \frac{1}{R_{k^*}}.$$

□

Effective ranks

Theorem 12. *Consider some positive summable sequence $\{\lambda_i\}_{i=1}^\infty$, and for any non-negative integer i denote*

$$r_i := \lambda_{i+1}^{-1} \sum_{j>i} \lambda_j.$$

Then $r_i > 1$ and $\sum_i r_i^{-1} = \infty$. Moreover, for any positive sequence $\{u_i\}$ such that $\sum_{i=0}^\infty u_i^{-1} = \infty$ and for every i $u_i > 1$, there exists a positive sequence $\{\lambda_i\}$ (unique up to constant multiplier) such that $r_i \equiv u_i$. The sequence is (a constant rescaling of)

$$\lambda_k = u_{k-1}^{-1} \prod_{i=0}^{k-2} (1 - u_i^{-1}).$$

Proof.

$$\sum_{i \geq k+1} \lambda_i = \sum_{i \geq k} \lambda_i - \lambda_k = (1 - r_{k-1}^{-1}) \sum_{i \geq k} \lambda_i.$$

Thus,

$$\sum_{i \geq k+1} \lambda_i = \prod_{i=0}^{k-1} \left(1 - r_i^{-1}\right) \cdot \sum_i \lambda_i,$$

which goes to zero if and only if $\sum_i r_i^{-1} = \infty$. On the other hand, we may rewrite the first equality in the proof as

$$\lambda_{k+1} r_k = \lambda_k r_{k-1}(1 - r_{k-1}^{-1}),$$

and hence

$$\lambda_k r_{k-1} = \prod_{i=0}^{k-2} \left(1 - r_i^{-1}\right) \lambda_1 r_0.$$

So for any sequence $\{u_i\}$ we can uniquely (up to a constant multiplier) recover the sequence $\{\lambda_i\}$ such that $r_i = u_i$ — the only candidate is

$$\lambda_k = u_{k-1}^{-1} \prod_{i=0}^{k-2}(1 - u_i^{-1}).$$

However, for such $\{\lambda_i\}$ one can compute

$$\sum_{i=1}^{k} \lambda_i = 1 - \prod_{i=0}^{k-1}(1 - u_i^{-1}),$$

so the resulting sequence $\{\lambda_i\}$ sums to 1, and

$$r_k = \lambda_{k+1}^{-1} \sum_{i>k} \lambda_i = \lambda_{k+1}^{-1} \prod_{i=0}^{k-1}(1 - u_i^{-1}) = u_k.$$

□

Benign sequences

Here we prove the following theorem.

Theorem 15. *Define $\lambda_{k,n} := \mu_k(\mathbf{\Sigma}_n)$ for all k, n.*

1. *If $\lambda_{k,n} = k^{-\alpha} \ln^{-\beta}(k+1)$, then $\mathbf{\Sigma}_n$ is benign if and only if $\alpha = 1$ and $\beta > 1$.*
2. *If $\lambda_{k,n} = k^{-(1+\alpha_n)}$, then $\mathbf{\Sigma}_n$ is benign if and only if $\omega(1/n) = \alpha_n = o(1)$.*
3. *If*

$$\lambda_{k,n} = \begin{cases} k^{-\alpha} & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

then $\mathbf{\Sigma}_n$ is benign if and only if either $0 < \alpha < 1$, $p_n = \omega(n)$ and $p_n = o\left(n^{1/(1-\alpha)}\right)$ or $\alpha = 1$, $p_n = e^{\omega(\sqrt{n})}$ and $p_n = e^{o(n)}$.

4. If

$$\lambda_{k,n} = \begin{cases} \gamma_k + \epsilon_n & \text{if } k \leq p_n, \\ 0 & \text{otherwise,} \end{cases}$$

and $\gamma_k = \Theta(\exp(-k/\tau))$, then $\mathbf{\Sigma}_n$ is benign if and only if $p_n = \omega(n)$ and $ne^{-o(n)} = \epsilon_n p_n = o(n)$.

We start the proof by proving two auxiliary lemmas.

Lemma 78. *Fix some sequence $(\lambda_i)_{i=1}^{\infty}$ and define $r_k = \lambda_{k+1}^{-1} \sum_{i>k} \lambda_i$ for any non-negative integer k. Suppose b is some constant, and $k^*(n) = \min\{k : r_k \geq bn\}$. Suppose also that the sequence $(r_n)_{n=1}^{\infty}$ is increasing. Then, as n goes to infinity, $k^*(n)/n$ goes to zero if and only if r_n/n goes to infinity.*

Proof. We prove the "if" part separately from the "only if" part.

1. **If $k^*(n)/n \to 0$ then $r_n/n \to \infty$.**

   Fix some $C > 1$. Since $k^*(n)/n \to 0$, there exists some N_C such that for any $n \geq N_C$, $k^*(n) < n/C$. Thus, for all $n > N_C$,

   $$k^*(\lfloor Cn \rfloor) \leq n,$$
   $$r_n \geq r_{k^*(\lfloor Cn \rfloor)} \geq b\lfloor Cn \rfloor.$$

   Since the constant C is arbitrary, r_n/n goes to infinity.

2. **If $r_n/n \to \infty$ then $k^*(n)/n \to 0$.**

   Fix some constant $C > 1$. Since $r_n/n \to \infty$ there exists some N_C such that for any $n \geq N_C$, $r_n > Cn$. Thus, for any $n > CN_C/b$

   $$r_{\lceil nb/C \rceil} \geq bn,$$
   $$k^*(n) \leq \lceil nb/C \rceil.$$

   Since the constant C is arbitrary, $k^*(n)/n$ goes to zero.

□

Lemma 79. *Suppose the sequence $\{r_i\}$ is increasing and $r_n/n \to \infty$ as $n \to \infty$. Then a sufficient condition for $\frac{n}{R_{k^*(n)}} \to 0$ is*

$$r_k^{-2} = o(r_k^{-1} - r_{k+1}^{-1}) \text{ as } k \to \infty.$$

For example, this condition holds for $r_n = n \log n$.

Proof. We need to show that

$$\frac{n}{R_{k^*(n)}} = \frac{n\sum_{i>k^*(n)}\lambda_i^2}{\left(\sum_{i>k^*(n)}\lambda_i\right)^2} = \frac{n\sum_{i>k^*(n)}\lambda_i^2}{\lambda_{k^*(n)+1}^2 r_{k^*(n)}^2} \to 0.$$

Since $r_{k^*(n)} \geq bn$ and $\lim_{n\to\infty} k^*(n) = \infty$, it is enough to prove that $\frac{\sum_{i>k}\lambda_i^2}{\lambda_{k+1}^2 r_k} \to 0$ as k goes to infinity. Since

$$\lambda_{k+2}r_{k+1} = \lambda_{k+1}r_k(1 - r_k^{-1}),$$

we can write that

$$\begin{aligned}\lambda_{k+1+l}r_{k+l} &= \lambda_{k+1}r_k \prod_{i=k}^{k+l-1}(1 - r_i^{-1}) \\ &\leq \lambda_{k+1}r_k \exp\left(-\sum_{i=k}^{k+l-1} r_i^{-1}\right)\end{aligned}$$

which yields

$$\frac{\lambda_{k+1+l}}{\lambda_{k+1}r_k} \leq r_{k+l}^{-1}\exp\left(-\sum_{i=k}^{k+l-1} r_i^{-1}\right).$$

Thus, we obtain

$$\frac{\sum_{i>k}\lambda_i^2}{\lambda_{k+1}^2 r_k} \leq r_k \sum_{i\geq k} r_i^{-2}\exp\left(-2\sum_{j=k}^{i-1} r_j^{-1}\right),$$

and it is sufficient to prove that the latter quantity goes to zero. We write

$$\begin{aligned}r_k \sum_{i\geq k} r_i^{-2}\exp\left(-2\sum_{j=k}^{i-1} r_j^{-1}\right) &= \frac{\sum_{i\geq k} r_i^{-2}\exp\left(-2\sum_{j=k}^{i-1} r_j^{-1}\right)}{r_k^{-1}} \\ &= \frac{\sum_{i\geq k} r_i^{-2}\exp\left(-2\sum_{j=0}^{i-1} r_j^{-1}\right)}{r_k^{-1}\exp\left(-2\sum_{j=0}^{k-1} r_j^{-1}\right)}.\end{aligned}$$

Since both numerator and denominator are decreasing in k and go to zero as $k \to \infty$, we can apply the Stolz–Cesáro theorem (an analog of L'Hôpital's rule for discrete sequences):

$$\begin{aligned}
\lim_{k\to\infty} \frac{\sum_{i\geq k} r_i^{-2} \exp\left(-2\sum_{j=0}^{i-1} r_j^{-1}\right)}{r_k^{-1}\exp\left(-2\sum_{j=0}^{k-1} r_j^{-1}\right)} &= \lim_{k\to\infty} \frac{r_k^{-2}\exp\left(-2\sum_{j=0}^{k-1} r_j^{-1}\right)}{(r_k^{-1} - e^{-2r_k^{-1}} r_{k+1}^{-1})\exp\left(-2\sum_{j=0}^{k-1} r_j^{-1}\right)} \\
&= \lim_{k\to\infty} \frac{r_k^{-2}}{(r_k^{-1} - e^{-2r_k^{-1}} r_{k+1}^{-1})} \\
&\qquad \text{(since, for large enough } k,\ e^{-2r_k^{-1}} \leq 1 - r_k^{-1}) \\
&\leq \lim_{k\to\infty} \frac{r_k^{-2}}{r_k^{-1} - r_{k+1}^{-1} + r_k^{-1} r_{k+1}^{-1}} \\
&= 0,
\end{aligned}$$

where the last line is due to our sufficient condition. □

Now we are ready to prove Theorem 15.

Part 1, if direction, first term: We have

$$r_0(\mathbf{\Sigma}_n) = \lambda_1^{-1}\sum_{i=1}^{\infty}\lambda_i = \sum_{i=1}^{\infty}\frac{\log^{\beta}(2)}{i\log^{\beta}(1+i)},$$

which is $O(1)$ for $\beta > 1$ since the function $f(x) = x^{-1}\left(\log(2)/\log(x)\right)^{\beta}$ has finite integral on $[1, +\infty)$.

Part 1, if direction, second term: By Lemma 78, it suffices to prove that $\lim_{n\to\infty}\frac{r_n}{n} = \infty$. This holds because

$$r_n = \frac{\sum_{i>n}\frac{1}{i\log^{\beta}(1+i)}}{\frac{1}{(n+1)\log^{\beta}(2+n)}} = \Theta(n\log n),$$

since $\beta > 1$.

Part 1, if direction, third term: By Lemma 79, it suffices to prove that $r_k^{-2} = o(r_k^{-1} - r_{k+1}^{-1})$, that is

$$\lim_{k\to\infty}\frac{r_k^{-2}}{r_k^{-1} - r_{k+1}^{-1}} = 0$$

or, equivalently,

$$\lim_{k\to\infty}\frac{r_{k+1}}{r_k(r_{k+1} - r_k)} = 0.$$

As argued above, when $\alpha = 1$ and $\beta > 1$, $r_k = \Theta(k \log k)$, so it suffices to show that $\lim_{k\to\infty}(r_{k+1} - r_k) = \infty$. We have

$$\begin{aligned} r_{k+1} - r_k &= \frac{\sum_{i>k+1} \lambda_i}{\lambda_{k+2}} - \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \\ &= \frac{\left((\lambda_{k+1} - \lambda_{k+2}) \sum_{i>k+1} \lambda_i\right) - \lambda_{k+1}\lambda_{k+2}}{\lambda_{k+1}\lambda_{k+2}} \\ &= \left(\left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \sum_{i>k+1} \lambda_i\right) - 1 \end{aligned}$$

so it suffices to show that

$$\lim_{k\to\infty} \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \sum_{i>k+1} \lambda_i = \infty.$$

Since λ_i is non-increasing, we have

$$\begin{aligned} \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \sum_{i>k+1} \lambda_i &\geq \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \int_{k+1}^{\infty} \frac{1}{x \log^{\beta} x} \, dx \\ &= \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \frac{1}{(\beta - 1)\log^{\beta-1}(k+1)} \\ &= \frac{(k+2)\log^{\beta}(k+3) - (k+1)\log^{\beta}(k+2)}{(\beta-1)\log^{\beta-1}(k+1)}. \end{aligned}$$

If we define f on the positive reals by $f(x) = x \log^{\beta}(x+1)$, then f is convex, and, since $f'(x) = \frac{\beta x \log^{\beta-1}(x+1)}{x+1} + \log^{\beta}(x+1)$, we have

$$\frac{(k+2)\log^{\beta}(k+3) - (k+1)\log^{\beta}(k+2)}{(\beta-1)\log^{\beta-1}(k+1)} \geq \frac{\frac{\beta(k+1)\log^{\beta-1}(k+2)}{k+2} + \log^{\beta}(k+2)}{(\beta-1)\log^{\beta-1}(k+1)},$$

which goes to infinity for large k, completing the proof of the "if" direction of the third term of Part 1.

Part 1, only if direction, $\alpha > 1$: If $\alpha > 1$, then

$$\begin{aligned} r_n &= \frac{\sum_{i>n} \frac{1}{i^a \log^{\beta}(1+i)}}{\frac{1}{n^a \log^{\beta}(1+n)}} \\ &\leq n^{\alpha} \sum_{i>n} \frac{\log^{\beta}(1+n)}{i^a \log^{\beta}(1+i)} \\ &\leq n^{\alpha} \sum_{i>n} \frac{1}{i^a} \\ &= n^{\alpha} O(n^{1-\alpha}), \end{aligned}$$

which does not grow faster than n. Thus, by Lemma 78, $k^*(n)/n$ does not go to zero.

Part 1, only if direction, $\alpha < 1$, or $\alpha = 1$ and $\beta \leq 1$: In this case, $\sum_{i=1}^{\infty} \frac{1}{i^\alpha \log^\beta(1+i)}$ diverges, so $\frac{r_0(\mathbf{\Sigma}_n)}{n}$ does not go to zero.

Before starting on Part 2, let us define $r_{k,n} = r_k(\mathbf{\Sigma}_n)$ and $R_{k,n} = R_k(\mathbf{\Sigma}_n)$.

Part 2, if direction, first term: We have

$$r_{0,n} = \sum_{i=1}^{\infty} \frac{1}{i^{1+\alpha_n}} \leq 1 + \frac{1}{\alpha_n},$$

so $\frac{r_{0,n}}{n} \leq \frac{1+\frac{1}{\alpha_n}}{n}$ which goes to zero with n if $\alpha_n = \omega(1/n)$.

Part 2, if direction, second term: First,

$$\begin{aligned} r_{k,n} &= (k+1)^{1+\alpha_n} \sum_{i>k} i^{-(1+\alpha_n)} \\ &\geq (k+1)^{1+\alpha_n} \int_{k+1}^{\infty} x^{-(1+\alpha_n)} dx \\ &= \frac{k+1}{\alpha_n}. \end{aligned}$$

Thus, $k^*(n) = O(\alpha_n n)$, so that $\frac{k^*(n)}{n} = O(\alpha_n) = o(1)$.

Part 2, if direction, third term: We bound $R_{k,n}$ from below by separately bounding its numerator and denominator:

$$\begin{aligned} \sum_{i>k} i^{-(1+\alpha_n)} &\geq \int_{k+1}^{\infty} x^{-(1+\alpha_n)}\ dx \\ &= \frac{1}{\alpha_n (k+1)^{\alpha_n}}, \end{aligned}$$

and

$$\begin{aligned} \sum_{i>k} i^{-2(1+\alpha_n)} &\leq \int_{k}^{\infty} x^{-2(1+\alpha_n)}\ dx \\ &= \frac{1}{k^{1+2\alpha_n}(2\alpha_n+1)}, \end{aligned}$$

so that

$$R_{k,n} \geq \frac{k^{1+2\alpha_n}(2\alpha_n+1)}{\alpha_n^2 (k+1)^{2\alpha_n}} \geq \frac{k}{\alpha_n^2} \times \left(1 - \frac{1}{k+1}\right)^{2\alpha_n}. \tag{A.6}$$

So now we want a lower bound on $k^*(n)$. For that, we need an upper bound on $r_{k,n}$, and

$$\begin{aligned} r_{k,n} &\le (k+1)^{1+\alpha_n} \int_k^\infty x^{-(1+\alpha_n)} dx \\ &= \frac{(k+1)}{\alpha_n} \times \left(1+\frac{1}{k}\right)^{\alpha_n} \\ &\le \frac{2k}{\alpha_n} e^{\alpha_n/k}. \end{aligned}$$

This implies $\frac{2k^*(n)}{\alpha_n} e^{\alpha_n/k^*(n)} \ge bn$. This, together with the fact that, for $u > 1$, $ue^{1/u}$ is an increasing function of u, implies that, for large enough n, $k^*(n) \ge \alpha_n bn/3$. Since $\alpha_n = \omega(1/n)$, this implies that $k^*(n) = \omega(1)$. Combining this with (A.6), for large enough n

$$R_{k^*(n),n} \ge \frac{k^*(n)}{\alpha_n^2} e^{-\alpha_n/k^*(n)} \ge \frac{k^*(n)}{2\alpha_n^2} \ge \frac{bn}{6\alpha_n}.$$

Thus $n/R_{k^*(n),n} = O(\alpha_n) = o(1)$.
Part 2, only if direction, $\alpha_n \ne \omega(1/n)$: We have

$$r_{0,n} = \sum_{i=1}^\infty \frac{1}{i^{1+\alpha_n}} \ge \frac{1}{\alpha_n},$$

so $\frac{r_{0,n}}{n} \ge \frac{1}{\alpha_n n}$, which does not go to zero unless $\alpha_n = \omega(1/n)$.
Part 2, only if direction, $\alpha_n \ne o(1)$: Recall that, in the proof of the "if" direction of the third term, we showed that $k^*(n) \ge \alpha_n bn/3$. This implies that $\frac{k^*(n)}{n} = \Omega(\alpha_n)$, so it is required to have $\alpha_n = o(1)$.
Part 3: Suppose that $\mathbf{\Sigma}_n$ is benign. Then because $r_k(\mathbf{\Sigma}_n) \le p_n - k$, we must have $p_n = \omega(n)$. Thus, we can restrict our attention to the sequences for which $p_n = \omega(n)$ and find the necessary and sufficient conditions for that class.

Next, for any positive α and any natural number $k \in [1, p_n)$, we can write

$$\int_k^{p_n} x^{-\alpha}\, dx \ge \sum_{i=k+1}^{p_n} i^{-\alpha} \ge \int_{k+1}^{p_n} x^{-\alpha}\, dx,$$

$$F(p_n) - F(k) \ge \sum_{i=k+1}^{p_n} i^{-\alpha} \ge F(p_n) - F(k+1),$$

where

$$F(x) = \begin{cases} \frac{1}{1-\alpha} x^{1-\alpha}, & \text{for } \alpha \ne 1, \\ \ln(x), & \text{for } \alpha = 1. \end{cases}$$

As the sequence can only be benign if $k^* = o(n)$, we can only consider values of k that do not exceed some constant fraction of n, e.g. $n/2$. Since $p_n = \omega(n)$, noting that, for $x > 0$, the sign of $\frac{1}{1-\alpha}x^{1-\alpha}$ flips when α crosses 1, we can write, uniformly for all $k \in [1, n/2]$,

$$\sum_{i=k+1}^{p_n} i^{-\alpha} = \begin{cases} \Theta_\alpha\left(p_n^{1-\alpha}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(\ln(p_n/k)\right), & \text{for } \alpha = 1, \\ \Theta_\alpha\left(k^{1-\alpha}\right), & \text{for } \alpha > 1. \end{cases}$$

Recall that we consider $\lambda_{i,n} = i^{-\alpha}$ for $i \le p_n$. Using the formula above, we get uniformly for all $k \in [1, n/2]$

$$r_k(\mathbf{\Sigma}_n) = \begin{cases} \Theta_\alpha\left(k^\alpha p_n^{1-\alpha}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(k\ln(p_n/k)\right), & \text{for } \alpha = 1, \\ \Theta_\alpha\left(k\right), & \text{for } \alpha > 1. \end{cases}$$

Recall that $k^* = \min\{k : r_k(\mathbf{\Sigma}_n) \ge bn\}$. We compute

$$k^* = \begin{cases} \Theta_\alpha\left(p_n^{1-\frac{1}{\alpha}} n^{\frac{1}{\alpha}}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(\frac{n}{\ln(p_n/n)}\right), & \text{for } \alpha = 1, \\ \Theta_\alpha\left(n\right), & \text{for } \alpha > 1. \end{cases}$$

One can see that for $\alpha > 1$, $k^* = \Omega_\alpha(n)$, so the sequence is not benign for $\alpha > 1$. On the other hand, $k^* = o(n)$ for $\alpha \le 1$.

Next, analogously to the asymptotics for $r_k(\mathbf{\Sigma})$, we have

$$r_k(\mathbf{\Sigma}_n^2) = \begin{cases} \Theta_\alpha\left(k^{2\alpha} p_n^{1-2\alpha}\right), & \text{for } \alpha \in (0,0.5), \\ \Theta_\alpha\left(k\ln(p_n/k)\right), & \text{for } \alpha = 0.5, \\ \Theta_\alpha\left(k\right), & \text{for } \alpha \in (0.5,1]. \end{cases}$$

Since $R_k = \frac{r_k(\mathbf{\Sigma})^2}{r_k(\mathbf{\Sigma}^2)}$, we can write uniformly for all $k \in [1, n/2]$

$$R_k = \begin{cases} \Theta_\alpha\left(p_n\right), & \text{for } \alpha \in (0,0.5), \\ \Theta_\alpha\left(\frac{p_n}{\ln(p_n/k)}\right), & \text{for } \alpha = 0.5, \\ \Theta_\alpha\left(k^{2\alpha-1} p_n^{2-2\alpha}\right), & \text{for } \alpha \in (0.5,1), \\ \Theta_\alpha\left(\ln(p_n/k)^2\right), & \text{for } \alpha = 1. \end{cases}$$

Now we plug in k^* instead of k. Recall that $p_n/k^* = \Theta_\alpha\left((p_n/n)^{1/\alpha}\right)$ for $\alpha \in (0,1)$, and $p_n/k^* = \Theta_\alpha\left(p_n/n\ln(p_n/n)\right)$ for $\alpha = 1$. We get

$$R_{k^*} = \begin{cases} \Theta_\alpha\left(p_n\right), & \text{for } \alpha \in (0,0.5), \\ \Theta_\alpha\left(n\frac{p_n/n}{\ln(p_n/n)}\right), & \text{for } \alpha = 0.5, \\ \Theta_\alpha\left(n\left(\frac{p_n}{n}\right)^{\frac{1}{\alpha}-1}\right), & \text{for } \alpha \in (0.5,1), \\ \Theta_\alpha\left(\ln(p_n/n)^2\right), & \text{for } \alpha = 1. \end{cases}$$

Since $p_n = \omega(n)$, for any $\alpha \in (0,1)$, $R_{k^*} = \omega(n)$. For $\alpha = 1$ the necessary and sufficient for $R_{k^*} = \omega(n)$ is $\ln(p_n/n) = \omega(\sqrt{n})$.

So far, we obtained the necessary and sufficient conditions for the last terms to go to zero. Now let's look at the upper bound for the first term: we write, for $\alpha \in (0,1]$,

$$r_0 = \sum_{i=1}^{p_n} i^{-\alpha} = \begin{cases} \Theta_\alpha\left(p_n^{1-\alpha}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(\ln p_n\right), & \text{for } \alpha = 1. \end{cases}$$

Thus, for $\alpha < 1$, $r_0(\mathbf{\Sigma}_n)/n$ goes to zero if and only if $p_n = o\left(n^{1/(1-\alpha)}\right)$, and for $\alpha = 1$, $r_0(\mathbf{\Sigma}_n)/n$ goes to zero if and only if $\ln(p_n) = o(n)$.

Part 4: Suppose that $\mathbf{\Sigma}_n$ is benign. Then because $r_k(\mathbf{\Sigma}_n) \leq p_n - k$, we must have $p_n = \omega(n)$. Also,

$$\begin{aligned} \operatorname{tr}(\mathbf{\Sigma}_n) &= \Theta\left(1 - e^{-p_n/\tau} + p_n\epsilon_n\right) \\ &= \Theta\left(1 + p_n\epsilon_n\right), \end{aligned}$$

and so $p_n\epsilon_n = o(n)$. Since $\mathbf{\Sigma}_n$ benign implies $k^* = o(n)$, and hence $k^* = o(p_n)$, we consider $k = o(p_n)$. In this regime,

$$\begin{aligned} \sum_{i>k} \lambda_i &= \Theta\left(e^{-k/\tau} - e^{-p_n/\tau} + (p_n - k)\epsilon_n\right) \\ &\leq \Theta\left(e^{-k/\tau} + p_n\epsilon_n\right). \end{aligned}$$

Thus, whenever $k \leq p_n$,

$$r_k(\mathbf{\Sigma}_n) \leq \Theta\left(\frac{e^{-k/\tau} + p_n\epsilon_n}{e^{-k/\tau} + \epsilon_n}\right).$$

Notice that

$$\frac{d}{dx}\frac{x + p_n\epsilon_n}{x + \epsilon_n} = \frac{\epsilon_n - p_n\epsilon_n}{(x+\epsilon_n)^2} < 0,$$

so k^* must be large enough to make

$$\frac{e^{-k/\tau} + p_n\epsilon_n}{e^{-k/\tau} + \epsilon_n} = \Omega(n).$$

Substituting $k = \tau \ln(n/(p_n\epsilon_n)) \pm O(1)$ gives

$$\begin{aligned} r_k(\mathbf{\Sigma}_n) &\leq \Theta\left(\frac{p_n\epsilon_n/n + p_n\epsilon_n}{p_n\epsilon_n/n + \epsilon_n}\right) \\ &= \Theta\left(\frac{p_n\epsilon_n}{p_n\epsilon_n/n}\right) \\ &= \Theta\left(n\right), \end{aligned}$$

which shows that $k^* \geq \tau \ln(n/(p_n\epsilon_n)) - O(1)$. Thus, if $\mathbf{\Sigma}_n$ is benign, we must have $k^* = o(n)$, that is, $\epsilon_n p_n = ne^{-o(n)}$.

Conversely, assume $p_n = \Omega(n)$ and $\epsilon_n p_n = ne^{-o(n)}$ (that is, $\ln(n/(p_n\epsilon_n)) = o(n)$). Set $k = \tau \ln(n/(p_n\epsilon_n)) - a$, for some a, which we shall see is $\Theta(1)$. Notice that $k = o(n)$, so $p_n - k = \Omega(p_n)$ and $e^{-p_n} = o(e^{-k})$. Thus,

$$\begin{aligned}
\sum_{i>k} \lambda_i &= \Theta\left(e^{-k/\tau} - e^{-p_n/\tau} + (p_n - k)\epsilon_n\right) \\
&= \Theta\left(e^{-k/\tau} + p_n\epsilon_n\right), \\
\sum_{i>k} \lambda_i^2 &= \Theta\left(e^{-2k/\tau} - e^{-2p_n} + (p_n - k)\epsilon_n^2\right) \\
&= \Theta\left(e^{-2k/\tau} + p_n\epsilon_n^2\right).
\end{aligned}$$

These imply

$$\begin{aligned}
\operatorname{tr}(\mathbf{\Sigma}_n) &= \Theta(1 + p_n\epsilon_n), \\
r_k(\mathbf{\Sigma}_n) &= \Theta\left(\frac{e^{-k/\tau} + p_n\epsilon_n}{e^{-k/\tau} + \epsilon_n}\right) \\
&= \Theta\left(\frac{ap_n\epsilon_n/n + p_n\epsilon_n}{ap_n\epsilon_n/n + \epsilon_n}\right) \\
&= \Theta\left(\frac{p_n\epsilon_n}{ap_n\epsilon_n/n}\right) \\
&= \Theta\left(n/a\right),
\end{aligned}$$

which shows that $k^* = \tau \ln(n/(p_n\epsilon_n)) + O(1)$. Also, we have

$$\begin{aligned}
R_k(\mathbf{\Sigma}_n) &= \Theta\left(\frac{\left(e^{-k/\tau} + p_n\epsilon_n\right)^2}{e^{-2k/\tau} + p_n\epsilon_n^2}\right) \\
&= \Theta\left(\frac{\left(p_n\epsilon_n/n + p_n\epsilon_n\right)^2}{p_n^2\epsilon_n^2/n^2 + p_n\epsilon_n^2}\right) \\
&= \Theta\left(\frac{p_n^2\epsilon_n^2}{p_n^2\epsilon_n^2/n^2 + p_n\epsilon_n^2}\right) \\
&= \Theta\left(\min\left\{n^2, p_n\right\}\right).
\end{aligned}$$

Now, it is clear that $p_n = \omega(n)$, $\epsilon_n p_n = o(n)$, and $\epsilon_n p_n = ne^{-o(n)}$ imply that $\mathbf{\Sigma}_n$ is benign.

A.3 Concentration inequalities

Lemma 80 (Mahalanobis norms of sub-Gaussian vectors)**.** *Suppose $\boldsymbol{z}$ is a σ-sub-Gaussian vector in $\mathbb{R}^p$. Consider $\mathbf{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ for some positive non-increasing sequence*

$\{\lambda_i\}_{i=1}^p$. *Then for some absolute constant c for any $t > 0$*

$$\mathbb{P}\left\{\|\mathbf{\Sigma}^{1/2}\boldsymbol{z}\|^2 > c\sigma^2\left(t\lambda_1 + \sum_i \lambda_i\right)\right\} \leq 2e^{-t/c}.$$

Proof. The argument consists of two parts: first, we obtain a bound that only works well in the case when all λ_i are approximately the same. Next, we split the sequence $\{\lambda_i\}$ into pieces with approximately equal values within each piece and obtain the final result by applying the first part of the argument to each piece.

First part: Consider a 1/4-net $\{\boldsymbol{u}_j\}_{j=1}^m$ on $\mathcal{S}^{p-1}$, such that $m \leq 9^p$. Note that for any vector $\boldsymbol{v} \in \mathcal{S}^{p-1}$ there exists an element $\boldsymbol{u}_j$ of that net such that $\langle \boldsymbol{v}, \boldsymbol{u}_j\rangle \geq 3/4 \cdot \|\boldsymbol{v}\|$. Thus, we have

$$\|\mathbf{\Sigma}^{1/2}\boldsymbol{z}\| \leq \frac{4}{3}\sqrt{\lambda_1}\max_j\langle \boldsymbol{z}, \boldsymbol{u}_j\rangle \leq 2\sqrt{\lambda_1}\max_j\langle \boldsymbol{z}, \boldsymbol{u}_j\rangle.$$

Since the random variable $\langle \boldsymbol{z}, \boldsymbol{u}_j\rangle$ is σ-sub-Gaussian, it also holds for any $t > 0$ and some absolute constant c that

$$\mathbb{P}(|\langle \boldsymbol{z}, \boldsymbol{u}_j\rangle| > t) \leq 2e^{-ct^2/\sigma^2},$$
$$\mathbb{P}(4\lambda_1\langle \boldsymbol{z}, \boldsymbol{u}_j\rangle^2 > 4\lambda_1 t\sigma^2) \leq 2e^{-ct}.$$

By multiplicity correction, we obtain

$$\mathbb{P}\left(\|\mathbf{\Sigma}^{1/2}\boldsymbol{z}\|^2 > 4\lambda_1\sigma^2 t + \frac{4\sigma^2\lambda_1\log 9}{c}p\right) \leq 2e^{-ct}.$$

We see that the random variable $\left(\|\mathbf{\Sigma}^{1/2}\boldsymbol{z}\|^2 - \frac{4\sigma^2\lambda_1\log 9}{c}p\right)_+$ has sub-Exponential norm bounded by $C\sigma^2\lambda_1$.

Second part: Now, instead of applying the result that we have just obtained to the whole vector $\boldsymbol{z}$, split it in the following way: define the sub-sequence $\{i_j\}$ in such that $i_1 = 1$, and for any $l \geq 1$ $i_{l+1} = \min\{i : \lambda_i < \lambda_{i_l}/2\}$. Denote $\boldsymbol{z}_l$ to be a sub-vector of $\boldsymbol{z}$ comprised of components from the i_l-th to $(i_{l+1}-1)$-th. Let $\mathbf{\Sigma}_l = \mathrm{diag}(\lambda_{i_l}, \dots, \lambda_{i_{l+1}-1})$.

Then by the initial argument, the random variable $\left(\|\mathbf{\Sigma}_l^{1/2}\boldsymbol{z}_l\|^2 - \frac{4\sigma^2\lambda_{i_l}\log 9}{c}(i_{l+1} - i_l)\right)_+$ has sub-Exponential norm bounded by $C\sigma^2\lambda_{i_l}$. Since each next λ_{i_l} is at most half of the previous, we obtain that the sum (over l) of those random variables has sub-Exponential norm at most $2C\sigma^2\lambda_1$. Combining this with the fact that

$$\sum_{i=i_l}^{i_{l+1}-1}\lambda_i \geq (i_{l+1} - i_l)\lambda_{i_{l+1}-1} \geq (i_{l+1} - i_l)\lambda_{i_{l+1}}/2,$$

we obtain that for some absolute constants $c_0, c_1, \dots$ for any $t > 0$

$$\begin{aligned} 2e^{-c_0 t} \geq & \mathbb{P}\left\{ \sum_l \left(\|\boldsymbol{\Sigma}_l^{1/2} \boldsymbol{z}_l\|^2 - c_1 \sigma^2 \lambda_{i_l} (i_{l+1} - i_l) \right) > c_2 \sigma^2 \lambda_1 t \right\} \\ \geq & \mathbb{P}\left\{ \|\boldsymbol{\Sigma}^{1/2} \boldsymbol{z}\|^2 \geq c_3 \sigma^2 \sum_i \lambda_i + c_2 \sigma^2 \lambda_1 t \right\}. \end{aligned}$$

□

Lemma 81 (Concentration of the sum of squared norms)**.** *Suppose $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ is a matrix with independent isotropic σ-sub-Gaussian rows $\boldsymbol{z}^1, \dots, \boldsymbol{z}^n$ (i.e. $\boldsymbol{Z}^\top = [\boldsymbol{z}^1, \dots, \boldsymbol{z}^n]$). Consider $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ for some positive non-increasing sequence $\{\lambda_i\}_{i=1}^p$. Then for some absolute constant c and any $t \in (0, n)$ with probability at least* $1 - 2\exp(-ct)$,

$$(n - \sqrt{nt}\sigma^2) \sum_{i>k} \lambda_i \leq \sum_{i=1}^n \|\boldsymbol{\Sigma}_{k:\infty}^{1/2} \boldsymbol{z}_{k:\infty}^i\|^2 \leq (n + \sqrt{nt}\sigma^2) \sum_{i>k} \lambda_i.$$

Proof. Since $\{\boldsymbol{z}_{k:\infty}^i\}_{i=1}^n$ are independent, isotropic and sub-Gaussian, $\|\boldsymbol{\Sigma}_{k:\infty}^{1/2} \boldsymbol{z}_{k:\infty}^i\|^2$ are independent sub-Exponential r.v.'s with expectation $\sum_{i>k} \lambda_i$ and sub-Exponential norms bounded by $c_1 \sigma^2 \sum_{i>k} \lambda_i$. Applying Bernstein's inequality gives

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n \|\boldsymbol{\Sigma}_{k:\infty}^{1/2} \boldsymbol{z}_{k:\infty}^i\|^2 - \sum_{i>k} \lambda_i \right| \geq t\sigma^2 \sum_{i>k} \lambda_i \right) \leq 2\exp\left(-c_2 \min(t, t^2) n \right).$$

Changing t to $\sqrt{t/n}$ gives the result. □

Lemma 82 (Weakened Hanson-Wright inequality)**.** *Suppose $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is a (random) PSD matrix and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a centered vector whose components $\{\varepsilon_i\}_{i=1}^n$ are independent and σ-sub-Gaussian. Then for some absolute constants c, C and any $t > 1$ with probability at least* $1 - 2e^{-t/c}$,

$$\boldsymbol{\varepsilon}^\top \boldsymbol{M} \boldsymbol{\varepsilon} \leq C t \sigma^2 tr(\boldsymbol{M}).$$

Proof. By Theorem 6.2.1 (Hanson-Wright inequality) in [55], for some absolute constant c_1 for any $t > 0$,

$$\mathbb{P}_{\boldsymbol{M}} \left\{ |\boldsymbol{\varepsilon}^\top \boldsymbol{M} \boldsymbol{\varepsilon} - \mathbb{E} \boldsymbol{\varepsilon}^\top \boldsymbol{M} \boldsymbol{\varepsilon}| \geq t \right\} \leq 2\exp\left(-c_1 \min\left\{ \frac{t^2}{\|\boldsymbol{M}\|_F^2 \sigma^4}, \frac{t}{\|\boldsymbol{M}\| \sigma^2} \right\} \right),$$

where $\mathbb{P}_{\boldsymbol{M}}$ denotes conditional probability given $\boldsymbol{M}$.

Since for any i, $\mathbb{E}\boldsymbol{\varepsilon}_i = 0$, and $\operatorname{Var}(\boldsymbol{\varepsilon}_i)$ is within a constant factor of σ^2, and since $\boldsymbol{M}$ is PSD, we have

$$\mathbb{E} \boldsymbol{\varepsilon}^\top \boldsymbol{M} \boldsymbol{\varepsilon} \leq c_2 \sigma^2 \operatorname{tr}(\boldsymbol{M}).$$

Moreover, since $\|\boldsymbol{M}\|_F^2 \le \operatorname{tr}(\boldsymbol{M})^2$ and $\|\boldsymbol{M}\| \le \operatorname{tr}(\boldsymbol{M})$, we obtain

$$\mathbb{P}_{\boldsymbol{M}}\left\{\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} > \sigma^2(c_2+t)\operatorname{tr}(\boldsymbol{M})\right\} \le 2\exp\{-c_1\min(t,t^2)\}.$$

Restricting to $t>1$ and adjusting the constants gives the result (note that since the RHS doesn't depend on $\boldsymbol{M}$, we can replace $\mathbb{P}_{\boldsymbol{M}}$ with $\mathbb{P}$). □

A.4 Controlling the singular values

In this section we use the following notation: for any matrix $\boldsymbol{M}$ we denote the element in the i-th row and the j-th column of $\boldsymbol{M}$ as $\boldsymbol{M}[i,j]$. We denote the i-th row of $\boldsymbol{M}$ as $\boldsymbol{M}[i,*]$ and the j-th column of $\boldsymbol{M}$ as $\boldsymbol{M}[*,j]$. For a vector $\boldsymbol{u}$ we denote it's i-th coordinate as $\boldsymbol{u}[i]$.

Lemma 83 (Bound on the norm of non-diagonal part of a Gram matrix)**.** *Denote $\mathring{\boldsymbol{A}}_k$ to be the matrix $\boldsymbol{A}_k$ with zeroed out diagonal elements: $\mathring{\boldsymbol{A}}_k[i,j] = (1-\delta_{i,j})\boldsymbol{A}_k[i,j]$. Then for some absolute constant c for any $t>0$ with probability at least $1-4e^{-t/c}$,*

$$\|\mathring{\boldsymbol{A}}_k\| \le c\sigma_x^2\sqrt{(t+n)\left(\lambda_{k+1}^2(t+n) + \sum_{i>k}\lambda_i^2\right)}.$$

Proof. We follow the lines of the decoupling argument from [56]. Consider a 1/4-net $\{\boldsymbol{u}_j\}_{j=1}^m$ on $\mathcal{S}^{n-1}$ s.t. $m \le 9^n$. Then

$$\|\mathring{\boldsymbol{A}}_k\| \le 2\max_j |\boldsymbol{u}_j^\top \mathring{\boldsymbol{A}}_k \boldsymbol{u}_j|.$$

Indeed, take $\boldsymbol{v} \in \mathcal{S}^{n-1}$ to be the eigenvector of $\mathring{\boldsymbol{A}}_k$ whose eigenvalue has the largest absolute value μ (i.e., $\|\mathring{\boldsymbol{A}}_k\| = \mu$), and let $\boldsymbol{u}_j$ be the closest point in the net to $\boldsymbol{v}$. Then

$$\begin{aligned}
\|\boldsymbol{v}-\boldsymbol{u}_j\| \le& 1/4,\\
\boldsymbol{u}_j^\top \boldsymbol{v} \ge& 3/4,\\
|\boldsymbol{u}_j^\top \mathring{\boldsymbol{A}}_k \boldsymbol{u}_j| \ge& |\boldsymbol{u}_j^\top \mathring{\boldsymbol{A}}_k \boldsymbol{v}| - |\boldsymbol{u}_j^\top \mathring{\boldsymbol{A}}_k(\boldsymbol{v}-\boldsymbol{u}_j)|\\
=& |\mu|\boldsymbol{u}_j^\top \boldsymbol{v} - |\boldsymbol{u}_j^\top \mathring{\boldsymbol{A}}_k(\boldsymbol{v}-\boldsymbol{u}_j)|\\
\ge& |\mu|\boldsymbol{u}_j^\top \boldsymbol{v} - \|\boldsymbol{u}_j\|\|\mathring{\boldsymbol{A}}_k\|\|\boldsymbol{v}-\boldsymbol{u}_j\|\\
\ge& |\mu|\left(\frac{3}{4}-\frac{1}{4}\right).
\end{aligned}$$

Denote the k-th coordinate of $\boldsymbol{u}_j$ as $\boldsymbol{u}_j[k]$. Note that

$$\boldsymbol{u}_j^\top \mathring{\boldsymbol{A}}_k \boldsymbol{u}_j = 4\mathbb{E}_T \sum_{k\in T\not\ni l} \boldsymbol{u}_j[k]\boldsymbol{u}_j[l]\mathring{\boldsymbol{A}}_k[k,l],$$

where the expectation is taken over a uniformly chosen random subset T of $\{1, \dots, n\}$ (since $\mathring{\boldsymbol{A}}_k$ has zeroed-out diagonal, we don't need to consider terms with $m = l$ which allows us to sum over $k \in T \not\ni l$). Thus,

$$\begin{aligned}
|\boldsymbol{u}_j^\top \mathring{\boldsymbol{A}}_k \boldsymbol{u}_j| &\le 4 \max_T \left| \sum_{l \in T \not\ni m} \boldsymbol{u}_j[l]\boldsymbol{u}_j[m]\mathring{\boldsymbol{A}}_k[l, m] \right| \\
&= 4 \max_T \left| \left\langle \sum_{l \in T} \boldsymbol{u}_j[l]\boldsymbol{X}_{k:\infty}[l, *], \sum_{m \notin T} \boldsymbol{u}_j[m]\boldsymbol{X}_{k:\infty}[m, *] \right\rangle \right|.
\end{aligned}$$

Fix j and denote

$$\begin{aligned}
\boldsymbol{\xi}^\top &:= \sum_{l \in T} \boldsymbol{u}_j[l]\boldsymbol{X}_{k:\infty}[l, *]\boldsymbol{\Sigma}_{k:\infty}^{-1/2}, \\
\boldsymbol{\eta}^\top &:= \sum_{m \notin T} \boldsymbol{u}_j[m]\boldsymbol{X}_{k:\infty}[m, *]\boldsymbol{\Sigma}_{k:\infty}^{-1/2}.
\end{aligned}$$

Note that since $\boldsymbol{u}_j$ is from the sphere, $\{\boldsymbol{X}_{k:\infty}[i, *]\}_{i=1}^n$ are independent, and l, m live in disjoint subsets, the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are independent sub-Gaussian with sub-Gaussian norms bounded by $C\sigma_x$ for some absolute constant C.

First, that means that for some absolute constant c_1 we have

$$\mathbb{P}\left\{ \left| \left\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{\xi}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{\eta} \right\rangle \right| \ge t\sigma_x \|\boldsymbol{\Sigma}\boldsymbol{\eta}\| \right\} \le 2e^{-c_1 t^2}.$$

Second, by Lemma 80, for some constant c_2 for any $t > 0$

$$\mathbb{P}\left\{ \|\boldsymbol{\Sigma}\boldsymbol{\eta}\|^2 \ge c_2 \sigma_x^2 \left(\lambda_{k+1}^2 t + \sum_{i>k} \lambda_i^2 \right) \right\} \le 2e^{-t/c_2}.$$

We obtain that for some absolute constant c for any $t > 0$ with probability at least $1 - 4e^{-t/c}$

$$\left| \left\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{\xi}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{\eta} \right\rangle \right| < c\sigma_x^2 \sqrt{t \left(\lambda_{k+1}^2 t + \sum_{i>k} \lambda_i^2 \right)}.$$

Finally, making multiplicity correction for all j (there are at most 9^n of them), and all subsets T (at most 2^n), we obtain that for some absolute constant c with probability at least $1 - 4e^{-t/c}$

$$\|\mathring{\boldsymbol{A}}_k\| \le c\sigma_x^2 \sqrt{(t+n) \left(\lambda_{k+1}^2 (t+n) + \sum_{i>k} \lambda_i^2 \right)}.$$

□

Lemma 84. *For some absolute constant c, for any $t > 0$, with probability at least $1 - 6e^{-t/c}$,*

$$\|\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\| \le c\sigma_x^2 \left(\lambda_{k+1}(t+n) + \sum_{i>k}\lambda_i\right).$$

Proof. Note that $\|\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\| \le \max_i \|\boldsymbol{X}_{i,*}\| + \|\mathring{\boldsymbol{A}}\|$. Combining Lemma 80 (with multiplicity correction) and Lemma 83 gives with probability $1 - 6e^{-t/c_1}$

$$\|\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\| \le c_1\sigma_x^2 \left((t + c_1\log n)\lambda_{k+1} + \sum_{i>k}\lambda_i + \sqrt{(t+n)\left(\lambda_{k+1}^2(t+n) + \sum_{i>k}\lambda_i^2\right)}\right).$$

Now note that

$$\begin{aligned}(t + c_1\log n)\lambda_{k+1} \le & c_1\sqrt{(t+n)\left(\lambda_{k+1}^2(t+n) + \sum_{i>k}\lambda_i^2\right)}\\ \le & c_1\sqrt{\lambda_{k+1}^2(t+n)^2 + \lambda_{k+1}(t+n)\sum_{i>k}\lambda_i}\\ \le & c_1\left(\lambda_{k+1}(t+n) + \sum_{i>k}\lambda_i\right),\end{aligned}$$

where we used $\sqrt{a^2+ab} \le a + b$ in the last transition. Removing the dominated (up to a constant multiplier) terms gives the result.

□

Lemma 18 (Controlling $\mu_1(\boldsymbol{A}_k)/\mu_n(\boldsymbol{A}_k)$ under sub-Gaussianity)**.** *For any $\gamma \in [0,1)$ and $\sigma_x > 0$ there exists $c > 0$ that only depends on σ_x and γ such that under Assumption NoncritReg(k,γ) the following holds: for any $L \ge 1$*

- *If $\rho_k \ge L^2$ and with probability at least $(1-\delta)^{1/n}$*

  $$\lambda + \|\boldsymbol{x}_{k:\infty}\|^2 \ge \frac{c}{L}\left(\lambda + \mathbb{E}\|\boldsymbol{x}_{k:\infty}\|^2\right),$$

  then with probability at least $1 - \delta - ce^{-n/c}$

  $$\mu_n(\boldsymbol{A}_k) \ge L^{-1}\mu_1(\boldsymbol{A}_k).$$

- *Suppose that it is known that with probability at least $ce^{-n/c}$ $\mu_n(\boldsymbol{A}_k) \ge L^{-1}\mu_1(\boldsymbol{A}_k)$. Then $\rho_k \ge \frac{1}{cL}$ and with probability at least $\left(1 - ce^{-n/c}\right)^{1/n}$*

  $$\lambda + \|\boldsymbol{x}_{k:\infty}\|^2 \ge \frac{1}{cL}\left(\lambda + \mathbb{E}\|\boldsymbol{x}_{k:\infty}\|^2\right).$$

Proof. We start with the high-probability bounds that we can derive assuming only sub-Gaussianity and independence of data vectors. By Lemma 80, for some absolute constant c and for any $t > 0$,

$$\mathbb{P}\left\{\|\boldsymbol{X}_{k:\infty}[i,*]\|^2 > c\sigma_x^2\left(t\lambda_{k+1} + \sum_{i>k}\lambda_i\right)\right\} \le 2e^{-t/c}.$$

By Lemma 83, for some absolute constant c and for any $t > 0$, with probability at least $1 - 4e^{-t/c}$,

$$\|\mathring{\boldsymbol{A}}_k\| \le c\sigma_x^2\sqrt{(t+n)\left(\lambda_{k+1}^2(t+n) + \sum_{i>k}\lambda_i^2\right)}.$$

Since $\|\boldsymbol{A}_k\| \le \lambda + \|\mathring{\boldsymbol{A}}_k\| + \max_i \|\boldsymbol{X}_{k:\infty}[i,*]\|$, the above two statements imply that for any $t > 0$ with probability at least $1 - 4e^{-n/c} - 2ne^{-t/c}$,

$$\begin{aligned}\mu_1(\boldsymbol{A}_k) \le& \lambda + c\sigma_x^2\sqrt{n\left(\lambda_{k+1}^2 n + \sum_{i>k}\lambda_i^2\right)} + c\sigma_x^2\left(t\lambda_{k+1} + \sum_{i>k}\lambda_i\right)\\ \le& \lambda + 2c\sigma_x^2\left((t+n)\lambda_{k+1} + \sum_{i>k}\lambda_i + \sqrt{n\sum_{i>k}\lambda_i^2}\right)\\ \le& \lambda + 3c\sigma_x^2\left((t+n)\lambda_{k+1} + \sum_{i>k}\lambda_i\right),\end{aligned}$$

where we used the following chain of inequalities to make the last transition:

$$2\sqrt{n\sum_{i>k}\lambda_i^2} \le 2\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} \le n\lambda_{k+1} + \sum_{i>k}\lambda_i.$$

On the same event,

$$\mu_n(\boldsymbol{A}_k) \ge \lambda + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2 - c\sigma_x^2\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

On the other hand, note that the sum of eigenvalues of $\boldsymbol{A}_k$ is equal to

$$\operatorname{tr}(\boldsymbol{A}_k) = \lambda n + \sum_{i=1}^n \|\boldsymbol{\Sigma}_{k:\infty}^{1/2}\boldsymbol{Z}_{k:\infty}[i,*]^\top\|^2.$$

By Lemma 81, for some absolute constant c and any $t \in (0,n)$, with probability at least $1 - 2e^{-ct}$,

$$(n - \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i \le \sum_{i=1}^n \|\boldsymbol{\Sigma}_{k:\infty}^{1/2}\boldsymbol{Z}_{k:\infty}[i,*]^\top\|^2 \le (n + \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i.$$

On this event

$$\begin{aligned}\mu_1(\boldsymbol{A}_k) \geq& \lambda + \left(1 - \sqrt{\frac{t}{n}}\sigma_x^2\right)\sum_{i>k}\lambda_i,\\ \mu_n(\boldsymbol{A}_k) \leq& \lambda + \left(1 + \sqrt{\frac{t}{n}}\sigma_x^2\right)\sum_{i>k}\lambda_i.\end{aligned}$$

Finally, note that $\mu_1(\boldsymbol{A}_k) \geq \lambda_{k+1}\|\boldsymbol{Z}_{k:\infty}[*,1]\|^2 + \lambda$. By Lemma 81, for some c_3 and for any $t \in (0, n)$, with probability. at least $1 - 2e^{-c_3 t}$,

$$\|\boldsymbol{Z}_{k:\infty}[*,1]\|^2 \geq n - \sqrt{nt}\sigma_x^2,$$

which means that

$$\mu_1(\boldsymbol{A}_k) \geq \lambda + n\lambda_{k+1}\left(1 - \sqrt{\frac{t}{n}}\sigma_x^2\right).$$

Combining all those bounds together gives that there is a constant c_x that only depends on σ_x such that with probability at least $1 - c_x e^{-n/c_x}$ all the following inequalities hold simultaneously:

$$\begin{aligned}\mu_1(\boldsymbol{A}_k) \leq& \lambda + c_x\left(n\lambda_{k+1} + \sum_{i>k}\lambda_i\right),\\ \mu_1(\boldsymbol{A}_k) \geq& \lambda + \frac{1}{c_x}\sum_{i>k}\lambda_i,\\ \mu_1(\boldsymbol{A}_k) \geq& \lambda + \frac{1}{c_x}n\lambda_{k+1},\\ \mu_n(\boldsymbol{A}_k) \geq& \lambda + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2 - c_x\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right),\\ \mu_n(\boldsymbol{A}_k) \leq& \lambda + c_x\sum_{i>k}\lambda_i,\\ \mu_n(\boldsymbol{A}_k) \leq& \lambda + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2.\end{aligned}$$

In view of the bounds that we derived above, the following inequality is a sufficient condition for the statement that with probability at least $1 - c_x e^{-n/c_x}$ the condition number of $\boldsymbol{A}_k$ does not exceed L:

$$\frac{1}{L}\left(\lambda + c_x\left(n\lambda_{k+1} + \sum_{i>k}\lambda_i\right)\right) \leq \lambda + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2 - c_x\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

Note that for any $\zeta > 0$

$$\sqrt{n\sum_{i>k}\lambda_i^2} < 2\sqrt{n\sum_{i>k}\lambda_i^2} \le 2\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} \le \zeta n\lambda_{k+1} + \zeta^{-1}\sum_{i>k}\lambda_i,$$

which implies that for any ζ the following is also a sufficient condition:

$$\lambda + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2 \ge \lambda L^{-1} + c_x(1 + L^{-1} + \zeta)n\lambda_{k+1} + c_x(L^{-1} + \zeta^{-1})\sum_{i>k}\lambda_i.$$

Recall that $\lambda > -\gamma\sum_{i>k}\lambda_i$, so

$$\sum_{i>k}\lambda_i \le \frac{1}{1-\gamma}\left(\lambda + \sum_{i>k}\lambda_i\right),$$

which allows us to upper bound the right-hand side of that condition. We write

$$\begin{aligned}
&\lambda L^{-1} + c_x(1 + L^{-1} + \zeta)n\lambda_{k+1} + c_x(L^{-1} + \zeta^{-1})\sum_{i>k}\lambda_i\\
\le& L^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right) + c_x(1 + L^{-1} + \zeta)\rho_k^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right) + \frac{c_x(L^{-1} + \zeta^{-1})}{1-\gamma}\left(\lambda + \sum_{i>k}\lambda_i\right)\\
=&\left(\lambda + \sum_{i>k}\lambda_i\right)\left(L^{-1}\left(1 + c_x\rho_k^{-1} + \frac{c_x}{1-\gamma}\right) + \rho_k^{-1}\left(c_x + c_x\zeta\right) + \frac{c_x\zeta^{-1}}{1-\gamma}\right).
\end{aligned}$$

Now take $\zeta = \rho_k^{1/2}$ and a constant c that is big enough depending on γ and c_x. Then if $\rho_k > L^2 > 1$ and with probability at least $1-\delta$,

$$\lambda + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2 \ge \frac{c}{L}\left(\lambda + \sum_{i>k}\lambda_i\right),$$

then with probability at least $1 - \delta - c_x e^{-n/c_x}$,

$$\mu_n(\boldsymbol{A}_k) \ge L^{-1}\mu_1(\boldsymbol{A}_k).$$

Note that since the rows of $\boldsymbol{X}_{k:\infty}$ are i.i.d., the first condition is equivalent to that with probability at least $(1-\delta)^{1/n}$

$$\lambda + \|\boldsymbol{X}_{k:\infty}[1,*]\|^2 \ge \frac{c}{L}\left(\lambda + \sum_{i>k}\lambda_i\right).$$

Now let's derive a necessary condition. Suppose it is known that with probability at least $c_x e^{-n/c_x}$ $\mu_n(\boldsymbol{A}_k) \geq L^{-1}\mu_1(\boldsymbol{A}_k)$. Then

$$\begin{aligned}\lambda + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2 \geq& \frac{1}{L}\left(\lambda + \frac{1}{c_x}\sum_{i>k}\lambda_i\right),\\ \lambda + c_x\sum_{i>k}\lambda_i \geq& \frac{1}{L}\left(\lambda + \frac{1}{c_x}n\lambda_{k+1}\right).\end{aligned}$$

For the first equation, we can write

$$\begin{aligned}\lambda + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2 \geq& \frac{1}{L}\left(\lambda + \frac{1}{c_x}\sum_{i>k}\lambda_i\right)\\ \lambda(1 - L^{-1} + L^{-1}c_x^{-1}) + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2 \geq& \frac{1}{Lc_x}\left(\lambda + \sum_{i>k}\lambda_i\right),\\ \lambda + \min_i \|\boldsymbol{X}_{k:\infty}[i,*]\|^2 \geq& \frac{1}{Lc_x(1 - L^{-1} + L^{-1}c_x^{-1})}\left(\lambda + \sum_{i>k}\lambda_i\right)\\ \geq& \frac{1}{Lc_x}\left(\lambda + \sum_{i>k}\lambda_i\right),\end{aligned}$$

where we used the fact that $c_x > 1$ and $L > 1$.

When it comes to the second equation, we write

$$\begin{aligned}\lambda + c_x\sum_{i>k}\lambda_i \geq& \frac{1}{L}\left(\lambda + \frac{1}{c_x}n\lambda_{k+1}\right),\\ (L-1)\lambda + c_x L\sum_{i>k}\lambda_i \geq& \frac{1}{c_x}n\lambda_{k+1} = \frac{1}{c_x}\rho_k^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right),\\ (L-1)\left(\lambda + \sum_{i>k}\lambda_i\right) + (c_xL - L + 1)\sum_{i>k}\lambda_i \geq& \frac{1}{c_x}\rho_k^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right)\\ \left(L - 1 + \frac{c_xL - L + 1}{1-\gamma}\right)\left(\lambda + \sum_{i>k}\lambda_i\right) \geq& \frac{1}{c_x}\rho_k^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right)\\ \rho_k \geq& c_x^{-1}\left(L - 1 + \frac{c_xL - L + 1}{1-\gamma}\right)^{-1} \geq c^{-1}L^{-1},\end{aligned}$$

where c is a large enough constant that only depends on γ and c_x. □

Lemma 85. *Suppose assumptions NoncritReg(k, γ) and CondNum(k, δ, L) are satisfied and $\gamma < 1$. Then for some absolute constant c for any $t \in (0, n)$ with probability at least $1 - \delta - 2e^{-ct}$*

$$\frac{1}{L}\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda + \sum_{i>k}\lambda_i\right) \le \mu_n(\boldsymbol{A}_k) \le \mu_1(\boldsymbol{A}_k) \le L\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda + \sum_{i>k}\lambda_i\right).$$

Moreover, if $\delta < 1 - 4e^{-ct}$ for some $t \in (0, n)$, then

$$\frac{\lambda + \sum_{i>k}\lambda_i}{n\lambda_{k+1}} \ge \frac{1 - \sigma_x^2\sqrt{t/n}}{L + \frac{\gamma}{1-\gamma} + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}(1-\gamma)}}.$$

Proof. First of all, note that the sum of eigenvalues of $\boldsymbol{A}_k$ is equal to

$$\operatorname{tr}(\boldsymbol{A}_k) = \lambda n + \sum_{i=1}^n \|\boldsymbol{\Sigma}_{k:\infty}^{1/2}\boldsymbol{Z}_{k:\infty}[i,*]^\top\|^2.$$

By Lemma 81 for some absolute constant c and any $t \in (0, n)$ with probability at least $1 - 2e^{-ct}$

$$(n - \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i \le \sum_{i=1}^n \|\boldsymbol{\Sigma}_{k:\infty}^{1/2}\boldsymbol{Z}_{k:\infty}[i,*]^\top\|^2 \le (n + \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i.$$

Now we know that with probability at least $1-\delta-2\exp(-c_2 t)$ the following two conditions hold:

$$\mu_1(\boldsymbol{A}_k) \le L\mu_n(\boldsymbol{A}_k),$$
$$n\lambda + (n - \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i \le \sum_{i=1}^n \mu_i(\boldsymbol{A}_k) \le n\lambda + (n + \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i.$$

The first line of the display above implies that

$$n\mu_1(\boldsymbol{A}_k)/L \le \sum_{i=1}^n \mu_i(\boldsymbol{A}_k) \le n\mu_n(\boldsymbol{A}_k)\cdot L$$

Thus, with probability at least $1 - \delta - 2\exp(-c_2 t)$,

$$\frac{\lambda}{L} + \frac{n - \sqrt{nt}\sigma_x^2}{nL}\sum_{i>k}\lambda_i \le \mu_n(\boldsymbol{A}_k) \le \mu_1(\boldsymbol{A}_k) \le \lambda L + \frac{(n + \sqrt{nt}\sigma_x^2)L}{n}\sum_{i>k}\lambda_i,$$

$$\frac{1}{L}\left(\lambda + \sum_i \lambda_i\right) - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}L}\sum_{i>k}\lambda_i \le \mu_n(\boldsymbol{A}_k) \le \mu_1(\boldsymbol{A}_k) \le L\left(\lambda + \sum_i \lambda_i\right) + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}}\sum_{i>k}\lambda_i.$$

Using the fact that $\sum_{i>k}\lambda_i \le \left(\lambda + \sum_{i>k}\lambda_i\right)/(1-\gamma)$, we obtain

$$\frac{1}{L}\left(\lambda + \sum_{i>k}\lambda_i\right)\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right) \le \mu_n(\boldsymbol{A}_k) \le \mu_1(\boldsymbol{A}_k) \le L\left(\lambda + \sum_{i>k}\lambda_i\right)\left(1 + \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right),$$

which gives the first assertion of the lemma.

Next, note that $\mu_1(\boldsymbol{A}_k) \ge \lambda_{k+1}\|\boldsymbol{Z}_{k:\infty}[*,1]\|^2 + \lambda$. By Lemma 81 for some c_3 for any $t \in (0,n)$ w.p. at least $1 - 2e^{-c_3 t}$, $\|\boldsymbol{Z}_{k:\infty}[*,1]\|^2 \ge n - \sqrt{nt}\sigma_x^2$, which means that if $1 - \delta - 2e^{-c_2 t} - 2e^{-c_3 t} > 0$ then with positive probability

$$\begin{aligned}
\lambda L + \frac{(n+\sqrt{nt}\sigma_x^2)L}{n}\sum_{i>k}\lambda_i \ge& \lambda_{k+1}(n - \sqrt{nt}\sigma_x^2) + \lambda,\\
\lambda(L-1) + \frac{(n+\sqrt{nt}\sigma_x^2)L}{n}\sum_{i>k}\lambda_i \ge& \lambda_{k+1}(n - \sqrt{nt}\sigma_x^2),\\
\left(\lambda + \sum_{i>k}\lambda_i\right)(L-1) + \left(1 + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}}\right)\sum_{i>k}\lambda_i \ge& \lambda_{k+1}(n - \sqrt{nt}\sigma_x^2),\\
\left(\lambda + \sum_{i>k}\lambda_i\right)\left(L + \frac{\gamma}{1-\gamma} + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}(1-\gamma)}\right) \ge& \lambda_{k+1}(n - \sqrt{nt}\sigma_x^2).
\end{aligned}$$

Taking $c_4 = \min(c_2, c_3)$ we see that if $\delta < 1 - 4e^{-c_4 t}$, then

$$\frac{\lambda + \sum_{i>k}\lambda_i}{n\lambda_{k+1}} \ge \frac{1 - \sigma_x^2\sqrt{t/n}}{L + \frac{\gamma}{1-\gamma} + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}(1-\gamma)}}.$$

□

Lemma 26 (k can be taken to be k^*)**.** *Fix any constants $\gamma \in [0,1)$, $b > 0$, $L > 0$. Denote*

$$k^* = \min\{k : \rho_k > b\}.$$

There exist constants c, L' that only depend on σ_x, γ, b, L s.t. the following holds: suppose assumptions NoncritReg(k, γ) and CondNum(k, δ, L) hold for some $k \in [k^*, n]$. Then assumptions NoncritReg(k^*, γ) and CondNum($k^*, \delta + ce^{-n/c}, L'$) hold too.

Proof. First, by Lemma 85 for any $t \in (0,n)$ with probability at least $1 - \delta - 2e^{-c_1 t}$,

$$\frac{1}{L}\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda + \sum_{i>k}\lambda_i\right) \le \mu_n(\boldsymbol{A}_k) \le \mu_n(\boldsymbol{A}_{k^*}).$$

Next, by Lemma 84 we know that with probability at least $1 - 6e^{-t/c_3}$,

$$\mu_1(\boldsymbol{A}_{k^*}) \le c_3\sigma_x^2 \left(\lambda_{k^*+1}(t+n) + \sum_{i>k^*} \lambda_i \right) + \lambda.$$

By definition of k^* and ρ_k

$$\lambda_{k^*+1} n = \rho_{k^*}^{-1} \left(\lambda + \sum_{i>k^*} \lambda_i \right) \le b^{-1} \left(\lambda + \sum_{i>k^*} \lambda_i \right).$$

Therefore,

$$\lambda + \sum_{i>k} \lambda_i = \lambda + \sum_{i>k^*} \lambda_i - \sum_{i=k^*+1}^{k} \lambda_i \ge \lambda + \sum_{i>k^*} \lambda_i - n\lambda_{k^*+1} \ge (1 - b^{-1}) \left(\lambda + \sum_{i>k^*} \lambda_i \right).$$

Moreover, since $\lambda > -\gamma \sum_{i>k^*} \lambda_i$,

$$\begin{aligned} \lambda \le& \lambda + \sum_{i>k^*} \lambda_i, \\ \sum_{i>k^*} \lambda_i \le& \frac{1}{1-\gamma} \left(\lambda + \sum_{i>k^*} \lambda_i \right) \\ \lambda_{k^*+1}(t+n) \le& b^{-1}(1 + t/n) \left(\lambda + \sum_{i>k^*} \lambda_i \right). \end{aligned}$$

Thus, with probability at least $1 - \delta - 8e^{-t/c_4}$

$$\begin{aligned} \mu_n(\boldsymbol{A}_{k^*}) \ge& \frac{1}{L} \left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)} \right) (1 - b^{-1}) \left(\lambda + \sum_{i>k^*} \lambda_i \right), \\ \mu_1(\boldsymbol{A}_{k^*}) \le& \left(c_3\sigma_x^2 \left(\frac{1}{1-\gamma} + \frac{1}{b} \left(1 + \frac{t}{n} \right) \right) + 1 \right) \left(\lambda + \sum_{i>k^*} \lambda_i \right). \end{aligned}$$

Taking c_5 large enough (depending on L, b, σ_x and γ) and plugging in $t = n/c_5$ gives the result for $c = \max(8, c_4c_5)$ and

$$L' = \left(c_3\sigma_x^2 \left(\frac{1}{1-\gamma} + \frac{1}{b} \left(1 + c_5^{-1} \right) \right) + 1 \right) \div \left(\frac{1}{L} \left(1 - \frac{\sigma_x^2}{\sqrt{c_5}(1-\gamma)} \right) (1 - b^{-1}) \right).$$

The derivation of *NoncritReg*(k^*, γ) is obvious: indeed, assumption *NoncritReg*(k, γ) states that

$$\lambda > -\gamma \sum_{i>k} \lambda_i.$$

Since $k^* \geq k$, $\sum_{i>k} \lambda_i \leq \sum_{i>k^*} \lambda_i$, so

$$\lambda > -\gamma \sum_{i>k^*} \lambda_i,$$

which is exactly assumption *NoncritReg*(k^*, γ). □

A.5 Lower bounds

We reuse a very convenient tool for proving lower bounds: Lemma 77 from Appendix A.2. We restate it below for convenience.

Lemma 77. *Suppose that $\{\eta_i\}_{i=1}^p$ is a sequence of non-negative random variables, and that $\{t_i\}_{i=1}^p$ is a sequence of non-negative real numbers (at least one of which is strictly positive) such that, for some $\delta \in (0,1)$ for any $i \leq p$ with probability at least $1-\delta$, $\eta_i > t_i$. Then with probability at least $1-2\delta$,*

$$\sum_{i=1}^n \eta_i \geq \frac{1}{2}\sum_{i=1}^p t_i.$$

It turns out to be quite straightforward to express bias and variance terms as sums of non-negative series. This lemma allows us to give a separate high probability lower bound for each term in the series to obtain the high probability lower bound for the whole sum.

Variance term

The argument for lower bounding the variance term is the same as in Appendix A.2. We repeat it here because the result there was stated in a different form and in the ridgeless setting only. Small changes are required to deal with possibly negative regularization.

Lemma 22 (Lower bound for the variance term)**.** *Fix any constant $\gamma \in [0,1)$. There exists a constant c that only depends on σ_x and γ s.t. for any $k < n/c$ under assumptions NoncritReg(k,γ) and IndepCoord w.p. at least $1 - ce^{-n/c}$*

$$V \geq \frac{1}{cn}\sum_{i=1} \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\}.$$

Proof. The variance term can be written as

$$V = \operatorname{tr}\left(\boldsymbol{\Sigma}\boldsymbol{X}^\top\boldsymbol{A}^{-2}\boldsymbol{X}\right) = \sum_{i=1}^\infty \frac{\lambda_i^2\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-2}\boldsymbol{z}_i}{(1+\lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2},$$

where $\boldsymbol{z}_i$ are columns of matrix $\boldsymbol{Z}$ (recall that $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$). Note that every term in this sum is non-negative, even if $\boldsymbol{A}_{-i}$ is not PSD. Denote $\boldsymbol{A}_{-i+}$ to be the PSD square root of $\boldsymbol{A}_{-i}^2$, i.e., the matrix with the same eigendecomposition as $\boldsymbol{A}_{-i}$, but with eigenvalues substituted by their absolute values. It immediately follows that

$$V \geq \sum_{i=1}^{\infty} \frac{\lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i}{(1 + \lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i+}^{-1} \boldsymbol{z}_i)^2},$$

By Cauchy-Schwartz we have

$$\|\boldsymbol{z}_i\|^2 \cdot \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-2} \boldsymbol{z}_i \geq (\boldsymbol{z}_i^\top \boldsymbol{A}_{-i+}^{-1} \boldsymbol{z}_i)^2.$$

Thus,

$$V \geq \sum_{i=1}^{\infty} \frac{1}{\|\boldsymbol{z}_i\|^2 \left(1 + (\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i+}^{-1} \boldsymbol{z}_i)^{-1}\right)^2}.$$

Now our goal is to lower-bound the largest eigenvalues of $\boldsymbol{A}_{-i+}^{-1}$. Let's write

$$\boldsymbol{A}_{-i} = \lambda \boldsymbol{I}_n + \sum_{j \neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top.$$

The idea is, as always, to separate the first k coordinates. Our initial goal is to bound the norm of $\sum_{j\neq i, j>k} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top$. Using Lemma 84, for some absolute constant c_1 and for any $t > 0$, with probability at least $1 - 6e^{-t/c_1}$,

$$\left\|\sum_{j\neq i, j>k} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top\right\| \leq \left\|\sum_{j>k} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top\right\| = \|\boldsymbol{X}_{k:\infty} \boldsymbol{X}_{k:\infty}^\top\| \leq c_1 \sigma_x^2 \left(\lambda_{k+1}(t+n) + \sum_{i>k} \lambda_i\right)$$

The matrix $\sum_{j\neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top$ is a correction to $\sum_{j\neq i, j>k} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top$ of rank at most k. Therefore, with probability at least $1 - 6e^{-t/c_1}$ the bottom $n-k$ eigenvalues of $\sum_{j\neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top$ lie in the segment from 0 to $c_1 \sigma_x^2 \left(\lambda_{k+1}(t+n) + \sum_{i>k} \lambda_i\right)$. The matrix $\boldsymbol{A}_{-i}$ has the same eigenvalues, but with λ added to each one, so on the same event all the eigenvalues of $\boldsymbol{A}_{-i}$ are from λ to $\lambda + c_1 \sigma_x^2 \left(\lambda_{k+1}(t+n) + \sum_{i>k} \lambda_i\right)$. We can write

$$\begin{aligned} &c_1 \sigma_x^2 \left(\lambda_{k+1}(t+n) + \sum_{i>k} \lambda_i\right) + \lambda \\ \leq &c_1 \sigma_x^2 \left(\lambda_{k+1}(t+n) + \frac{1}{1-\gamma}\left(\lambda + \sum_{i>k} \lambda_i\right)\right) + \frac{\gamma}{1-\gamma}\left(\lambda + \sum_{i>k} \lambda_i\right), \end{aligned}$$

where we used that $\lambda > -\gamma \sum_{i>k} \lambda_i$ in the second line (for $\lambda < 0$ it implies $|\lambda| < \gamma \sum_{i>k} \lambda_i$). Moreover, for the left end of the segment we also have that either $\lambda > 0$ or

$$|\lambda| \leq \gamma \sum_{i>k} \lambda_i \leq \frac{\gamma}{1-\gamma}\left(\lambda + \sum_{i>k} \lambda_i\right).$$

Thus, for some constant c_2 which only depends on σ and γ, for any i with probability at least $1-6e^{-n/c_2}$, for any $j>k$

$$|\mu_j(\boldsymbol{A}_i)| \le c_2\left(\lambda_{k+1}n+\lambda+\sum_{i>k}\lambda_i\right).$$

In words, with high probability the matrix $\boldsymbol{A}_{-i}$ has at least $n-k$ eigenvalues whose magnitude is bounded by $c_2\left(\lambda_{k+1}n+\lambda+\sum_{i>k}\lambda_i\right)$. Recall that $\boldsymbol{A}_{-i+}$ is PSD with the same magnitudes of the eienvalues. Denote $\boldsymbol{P}_{i,k}$ to be the projector on the linear space spanned by the first k eigenvectors of $\boldsymbol{A}_{-i+}$. We can now write that with probability at least $1-6e^{-n/c_2}$

$$\boldsymbol{z}_i^\top \boldsymbol{A}_{-i+}^{-1}\boldsymbol{z}_i \ge \|(I-\boldsymbol{P}_{i,k})\boldsymbol{z}_i\|^2 c_2^{-1}\left(\lambda_{k+1}n+\lambda+\sum_{i>k}\lambda_i\right)^{-1}$$

Since $\boldsymbol{z}_i$ is independent of $\boldsymbol{P}_{i,k}$, by Theorem 6.2.1 (Hanson-Wright inequality) in [55], for some absolute constant c_2 and for any $t>0$,

$$\mathbb{P}\left\{\left|\|\boldsymbol{P}_{i,k}\boldsymbol{z}_i\|^2-\mathbb{E}_{\boldsymbol{z}_i}\|\boldsymbol{P}_{i,k}\boldsymbol{z}_i\|^2\right|\ge t\right\} \le 2\exp\left(-c_2^{-1}\min\left\{\frac{t^2}{\sigma_x^4\|\boldsymbol{P}_{i,k}^2\|_F^2},\frac{t}{\sigma_x^2\|\boldsymbol{P}_{i,k}^2\|}\right\}\right).$$

We have $\|\boldsymbol{P}_{i,k}^2\|_F^2=k$, $\|\boldsymbol{P}_{i,k}^2\|=1$, and $\mathbb{E}_{\boldsymbol{z}_i}\|\boldsymbol{P}_{i,k}\boldsymbol{z}_i\|^2=\operatorname{tr}(\boldsymbol{P}_{i,k})=k$ since $\boldsymbol{P}_{i,k}$ is an orthogonal projector of rank kThus, w.p. at least $1-2e^{-t/c_2}$,

$$\left|\|\boldsymbol{P}_{i,k}\boldsymbol{z}_i\|^2-k\right| \le \sigma_x^2\max(\sqrt{kt},t)\le(t+\sqrt{kt})\sigma_x^2.$$

Next, by Lemma 81 for some constant c_3 and any $t\in(0,n)$ w.p. at least $1-2e^{-t/c_3}$,

$$n-\sqrt{nt}\sigma_x^2\le\|\boldsymbol{z}_i\|^2\le n+\sqrt{nt}\sigma_x^2.$$

Take constant c_4 large enough depending on σ_x and set $t=n/c_4$. Then for any $k<n/c_5$, w.p. at least $1-10e^{-n/c_6}-\delta$,

$$\boldsymbol{z}_i^\top \boldsymbol{A}_{-i+}^{-1}\boldsymbol{z}_i \ge \frac{n}{c_7\left(\lambda_{k+1}n+\lambda+\sum_{i>k}\lambda_i\right)},$$

where constants c_5 and c_6 depend only on σ_x and constant c_7 depends only on σ_x and γ.

Rewrite this equation as

$$(\boldsymbol{z}_i^\top \boldsymbol{A}_{-i+}^{-1}\boldsymbol{z}_i)^{-1}\le c_7\left(\lambda_{k+1}+\frac{1}{n}\left(\lambda+\sum_{i>k}\lambda_i\right)\right)=c_7\lambda_{k+1}(\rho_k+1),$$

where $\rho_k:=\frac{1}{n\lambda_{k+1}}\left(\lambda+\sum_{i>k}\lambda_i\right)$.

On the same event

$$\frac{1}{\|\boldsymbol{z}_i\|^2\left(1+(\lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}_{-i+}^{-1}\boldsymbol{z}_i)^{-1}\right)^2}\geq\frac{1}{c_8 n\left(1+\frac{\lambda_{k+1}}{\lambda_i}(\rho_k+1)\right)^2},$$

where c_8 depends only on σ_x and γ.

Finally, by Lemma 77, we can convert lower bounds for separate non-negative terms into a lower bound on their sum: with probability at least $1-20e^{-n/c_6}$,

$$V\geq\frac{1}{8c_8 n}\sum_{i=1}^{p}\min\left\{1,\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\},$$

where we also used that $1/(a+b)^2\geq\min(a^{-2},b^{-2})/4$ for non-negative a,b. □

Bias term

Lemma 23 (Lower bound for the bias term)**.** *Fix any constant $L>0$. There exists c that only depends on σ_x and L s.t. for any $k\in\{1,2,\dots,p\}$ under assumptions PriorSigns($\bar{\boldsymbol{\theta}}$) and StableLowEig($k,\delta,L$) w.p. at least $1-2\delta-ce^{-n/c}$*

$$\mathbb{E}_{\boldsymbol{\theta}^*}B\geq\frac{1}{c}\sum_i\frac{\lambda_i\bar{\boldsymbol{\theta}}_i^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2},$$

where $\mathbb{E}_{\boldsymbol{\theta}^*}$ denotes the expectation over the random draw of $\boldsymbol{\theta}^*$ from the prior distribution described in assumption PriorSigns($\bar{\boldsymbol{\theta}}$).

Proof. Applying Sherman-Morrison-Woodbury yields

$$\left(\lambda\boldsymbol{I}_p+\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}=\lambda^{-1}\boldsymbol{I}_p-\lambda^{-2}\boldsymbol{X}^\top(\boldsymbol{I}_n+\lambda^{-1}\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}.$$

So,

$$\begin{aligned}\left(\lambda\boldsymbol{I}_p+\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{I}_p=&\left(\lambda\boldsymbol{I}_p+\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}(\lambda\boldsymbol{I}_p+\boldsymbol{X}^\top\boldsymbol{X}-\lambda\boldsymbol{I}_p)-\boldsymbol{I}_p\\=&-\lambda\left(\lambda\boldsymbol{I}_p+\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\\=&\boldsymbol{I}_p-\lambda^{-1}\boldsymbol{X}^\top(\boldsymbol{I}_n+\lambda^{-1}\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\\=&\boldsymbol{I}_p-\boldsymbol{X}^\top(\lambda\boldsymbol{I}_n+\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}.\end{aligned}$$

Thus, the bias term becomes

$$(\boldsymbol{\theta}^*)^\top\left(\boldsymbol{I}_p-\boldsymbol{X}^\top(\lambda\boldsymbol{I}_n+\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\right)\boldsymbol{\Sigma}\left(\boldsymbol{I}_p-\boldsymbol{X}^\top(\lambda\boldsymbol{I}_n+\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\right)\boldsymbol{\theta}_*$$

and taking expectation over the prior kills all the off-diagonal elements, so

$$\mathbb{E}_{\boldsymbol{\theta}^*}\mathcal{B} = \sum_i \left(\left(\boldsymbol{I}_p - \boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\right)\boldsymbol{\Sigma}\left(\boldsymbol{I}_p - \boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\right)\right)[i,i]\cdot \bar{\boldsymbol{\theta}}_i^2.$$

Let's compute the diagonal elements of the matrix

$$\left(\boldsymbol{I}_p - \boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\right)\boldsymbol{\Sigma}\left(\boldsymbol{I}_p - \boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\right).$$

The i-th diagonal element is equal to the bias term for the case when $\boldsymbol{\theta}^* = e_i$ — the i-th vector of the standard orthonormal basis. Note that the i-th row of $\boldsymbol{I}_p - \boldsymbol{X}^\top(\lambda\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}$ is equal to $e_i - \sqrt{\lambda_i}\boldsymbol{z}_i^\top(\lambda\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}$, so the i-th diagonal element of the initial matrix is given by

$$\begin{aligned}&\sum_{j=1}^p \lambda_i\left(e_i[j] - \sqrt{\lambda_i\lambda_j}\boldsymbol{z}_i^\top(\lambda\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{z}_j\right)^2\\ &\lambda_i\left(1 - \lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}^{-1}\boldsymbol{z}_i\right)^2 + \sum_{j\neq i}\lambda_i\lambda_j^2(\boldsymbol{z}_i^\top\boldsymbol{A}^{-1}\boldsymbol{z}_j)^2.\end{aligned}$$

Recall that $\boldsymbol{A} = \lambda\boldsymbol{I}_n + \sum_{i=0}^p\lambda_i\boldsymbol{z}_i\boldsymbol{z}_i^\top$, $\boldsymbol{A}_{-i} := \boldsymbol{A} - \lambda_i\boldsymbol{z}_i\boldsymbol{z}_i^\top$.

First, let's use Sherman-Morrison identity to convert $\boldsymbol{A}$ in $\boldsymbol{z}_i^\top\boldsymbol{A}^{-1}\boldsymbol{z}_i$ into $\boldsymbol{A}_{-i}$:

$$\begin{aligned}1 - \lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}^{-1}\boldsymbol{z}_i =& 1 - \lambda_i\boldsymbol{z}_i^\top\left(\boldsymbol{A}_{-i} + \lambda_i\boldsymbol{z}_i\boldsymbol{z}_i^\top\right)^{-1}\boldsymbol{z}_i\\ =& 1 - \lambda_i\boldsymbol{z}_i^\top\left(\boldsymbol{A}_{-i}^{-1} - \lambda_i\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i(1 + \boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^{-1}\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\right)\boldsymbol{z}_i\\ =& 1 - \lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i + \frac{\left(\lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i\right)^2}{1 + \lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i}\\ =& \frac{1}{1 + \lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i}.\end{aligned}$$

So the diagonal element becomes

$$\frac{\lambda_i}{(1 + \lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2} + \sum_{j\neq i}\lambda_i\lambda_j^2(\boldsymbol{z}_i^\top\boldsymbol{A}^{-1}\boldsymbol{z}_j)^2 \geq \frac{\lambda_i}{(1 + \lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2},$$

and thus

$$\mathbb{E}_{\boldsymbol{\theta}^*}B \geq \sum_i\frac{\lambda_i\bar{\boldsymbol{\theta}}_i^2}{(1 + \lambda_i\boldsymbol{z}_i^\top\boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2}.$$

Let's bound each term in that sum from below with high probability. By our assumptions, for any i with probability at least $1-\delta$

$$\mu_n(\boldsymbol{A}_{-i}) \geq \frac{1}{L}\left(\lambda + \sum_{j>k}\lambda_j\right).$$

Next,

$$\frac{\lambda_i}{(1+\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2} \geq \frac{\lambda_i}{\left(1+\lambda_i\mu_n(\boldsymbol{A}_{-i})^{-1}\|\boldsymbol{z}_i\|^2\right)^2},$$

and by Lemma 81 for some absolute constant c_1 for any $t \in (0, n)$ w.p. at least $1-2e^{-t/c_1}$ we have $\|\boldsymbol{z}_i\|^2 \leq n - \sqrt{tn}\sigma_x^2 \leq n/2$, where the last transition is true if additionally $t \leq n/(4\sigma_x^4)$.

Recall that $\rho_k := \frac{\lambda + \sum_{j>k}\lambda_j}{n\lambda_{k+1}}$. We obtain by plugging $t = n/(4\sigma_x^4)$ that w.p. at least $1-\delta-2e^{-n/c_2}$,

$$\frac{\lambda_i\bar{\boldsymbol{\theta}}_i^2}{(1+\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2} \geq \frac{\lambda_i\bar{\boldsymbol{\theta}}_i^2}{\left(1+\frac{L\lambda_i}{2\lambda_{k+1}\rho_k}\right)^2},$$

where c_2 only depends on σ_x.

Finally, since all the terms are non-negative and we need to obtain a lower bound on their sum, Lemma 77 gives the result.

□

Lemma 24. *For any $\gamma < 1$ there exists a constant c that only depends on γ and σ_x such that if assumptions CondNum(k, δ, L), NoncritReg(k, γ) and ExchCoord are satisfied for some $L \geq 1$ and $k \in \{1, 2, \dots, p\}$, then StableLowEig$(k, \delta + 2e^{-n/c}, cL)$ is also satisfied.*

Proof. First of all, note that Assumption *NoncritReg*(k, γ) with $\gamma < 1$ directly implies that $\lambda + \sum_{i>k}\lambda_i \geq 0$, which is the second part of Assumption *StableLowEig*(k, δ, L).

Next, by Lemma 85 for some absolute constant c_1 for any $t \in (0, n)$ with probability at least $1-\delta-2e^{-ct}$

$$\mu_n(\boldsymbol{A}_k) \geq \frac{1}{L}\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda + \sum_{i>k}\lambda_i\right).$$

Taking $t = n/c_2$ where c_2 is large enough depending on γ, σ_x we get that for c large enough with probability at least $1-2e^{-n/c}$

$$\mu_n(\boldsymbol{A}_k) \geq \frac{1}{cL}\left(\lambda + \sum_{i>k}\lambda_i\right).$$

Now we just need to propagate that result to $\boldsymbol{A}_{-i}$ for all i.

For $i \leq k$, we simply have $\boldsymbol{A}_{-i} \succeq \boldsymbol{A}_k$ (that is,$\boldsymbol{A}_{-i}$ is at least as large as $\boldsymbol{A}_k$ in the sense of Loewner order) with probability 1, so indeed $\forall i \leq k$

$$\mathbb{P}\left(\mu_n(\boldsymbol{A}_{-i}) \geq \frac{1}{cL}\left(\lambda + \sum_{i>k}\lambda_i\right)\right) \geq \mathbb{P}\left(\mu_n(\boldsymbol{A}_k) \geq \frac{1}{cL}\left(\lambda + \sum_{i>k}\lambda_i\right)\right) \geq 1-2e^{-n/c}.$$

When it comes to $i > k$, we can write

$$\begin{aligned}
\boldsymbol{A}_{-i} =& \lambda \boldsymbol{I}_n + \sum_{j \neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top \\
=& \lambda \boldsymbol{I}_n + \sum_{j \leq k} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top + \sum_{j > k, j \neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top \\
\succeq& \lambda \boldsymbol{I}_n + \lambda_1 \boldsymbol{z}_1 \boldsymbol{z}_1^\top + \sum_{j > k, j \neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top \\
\succeq& \lambda \boldsymbol{I}_n + \lambda_i \boldsymbol{z}_1 \boldsymbol{z}_1^\top + \sum_{j > k, j \neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top .
\end{aligned}$$

Now note that due to Assumption *ExchCoord*, the distribution of the matrix $\lambda \boldsymbol{I}_n + \lambda_i \boldsymbol{z}_1 \boldsymbol{z}_1^\top + \sum_{j>k, j\neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top$ is the same as the distribution of $\boldsymbol{A}_k = \lambda \boldsymbol{I}_n + \sum_{j>k} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top$. Therefore

$$\begin{aligned}
&\mathbb{P}\left(\mu_n(\boldsymbol{A}_{-i}) \geq \frac{1}{cL}\left(\lambda + \sum_{i>k} \lambda_i\right)\right) \\
\geq& \mathbb{P}\left(\mu_n\left(\lambda \boldsymbol{I}_n + \lambda_i \boldsymbol{z}_1 \boldsymbol{z}_1^\top + \sum_{j>k, j\neq i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top\right) \geq \frac{1}{cL}\left(\lambda + \sum_{i>k} \lambda_i\right)\right) \\
=& \mathbb{P}\left(\mu_n(\boldsymbol{A}_k) \geq \frac{1}{cL}\left(\lambda + \sum_{i>k} \lambda_i\right)\right) \\
\geq& 1 - 2e^{-n/c},
\end{aligned}$$

which finishes the proof. □

A.6 Deriving a useful identity

Motivated by the results of Section 2.3, we split the principal directions of the covariance matrix into two parts: small dimensional and high dimensional. The main idea of our argument is to use classical machinery (like some sort of uniform convergence argument) in the small dimensional subspace. To do this we write $\hat{\boldsymbol{\theta}}(\boldsymbol{y})^\top = \left[\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}^\top, \hat{\boldsymbol{\theta}}(\boldsymbol{y})_{k:\infty}^\top\right]$ and mentally split the search process for $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ into two parts: first, for any fixed $\boldsymbol{\theta}_{0:k}$, optimize for $\boldsymbol{\theta}_{k:\infty}$. Then only the first k coordinates are left. The result of that optimization in $\boldsymbol{\theta}_{k:\infty}$ is the following identity:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} + \boldsymbol{X}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{X}_{0:k} \hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} = \boldsymbol{X}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{y}. \tag{A.7}$$

The goal of this section is to derive this identity.

Derivation in the ridgeless case

In the ridgeless case we are simply dealing with projections, and $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ is the minimum norm interpolating solution. Note that $\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{k:\infty}$ is also the minimum norm solution to the equation $\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}_{k:\infty} = \boldsymbol{y} - \boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}$, where $\boldsymbol{\theta}_{k:\infty}$ is the variable. Thus, we can write

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{k:\infty} = \boldsymbol{X}_{k:\infty}^\top \left(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\right)^{-1} \left(\boldsymbol{y} - \boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}\right).$$

Now we need to minimize the norm in $\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}$ (our choice of $\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{k:\infty}$ already makes the solution interpolating): we need to minimize the norm of the following vector:

$$\boldsymbol{v}(\boldsymbol{\theta}_{0:k}) = \left[\boldsymbol{\theta}_{0:k}^\top, (\boldsymbol{y} - \boldsymbol{X}_{0:k}\boldsymbol{\theta}_{0:k})^\top \left(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\right)^{-1} \boldsymbol{X}_{k:\infty}\right]$$

As $\boldsymbol{\theta}_{0:k}$ varies, this vector sweeps an affine subspace of our Hilbert space. The vector $\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}$ gives the minimum norm if and only if for any additional vector $\boldsymbol{\eta}_{0:k}$ we have $\boldsymbol{v}(\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}) \perp \boldsymbol{v}(\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} + \boldsymbol{\eta}_{0:k}) - \boldsymbol{v}(\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k})$. Let's write out the second vector: $\forall \boldsymbol{\eta}_{0:k} \in \mathbb{R}^k$

$$\boldsymbol{v}(\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} + \boldsymbol{\eta}_{0:k}) - \boldsymbol{v}(\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}) = \left[\boldsymbol{\eta}_{0:k}^\top, -\boldsymbol{\eta}_{0:k}^\top\boldsymbol{X}_{0:k}^\top \left(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\right)^{-1} \boldsymbol{X}_{k:\infty}\right]$$

We see that the above mentioned orthogonality for any $\boldsymbol{\eta}_{0:k}$ is equivalent to the following:

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}^\top - \left(\boldsymbol{y} - \boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}\right)^\top \left(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\right)^{-1} \boldsymbol{X}_{0:k} = 0,& \\ \hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(y)_{0:k} = \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{y},&\end{aligned}$$

where we replaced $\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top =: \boldsymbol{A}_k$.

Checking for the case of non-vanishing regularization

So, now we have $\lambda \neq 0$ and we want to prove that $\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} = \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{y}$. Recall that

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\boldsymbol{y}) =& \boldsymbol{X}^\top(\lambda\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}, \\ \hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} =& \boldsymbol{X}_{0:k}^\top(\boldsymbol{A}_k + \boldsymbol{X}_{0:k}\boldsymbol{X}_{0:k}^\top)^{-1}\boldsymbol{y}.\end{aligned}$$

This identity yields

$$\begin{aligned}&\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k} \\ =&\boldsymbol{X}_{0:k}^\top(\boldsymbol{A}_k + \boldsymbol{X}_{0:k}\boldsymbol{X}_{0:k}^\top)^{-1}\boldsymbol{y} + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\boldsymbol{X}_{0:k}^\top(\boldsymbol{A}_k + \boldsymbol{X}_{0:k}\boldsymbol{X}_{0:k}^\top)^{-1}\boldsymbol{y} \\ =&\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}(\boldsymbol{A}_k + \boldsymbol{X}_{0:k}\boldsymbol{X}_{0:k}^\top)(\boldsymbol{A}_k + \boldsymbol{X}_{0:k}\boldsymbol{X}_{0:k}^\top)^{-1}\boldsymbol{y} \\ =&\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{y}.\end{aligned}$$

A.7 Variance

Recall that the variance term is

$$V = \frac{1}{v_{\boldsymbol{\varepsilon}}^2}\mathbb{E}_{\boldsymbol{\varepsilon}}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\|_{\boldsymbol{\Sigma}}^2 = \frac{1}{v_{\boldsymbol{\varepsilon}}^2}\mathbb{E}_{\boldsymbol{\varepsilon}}\|\boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\varepsilon}\|_{\boldsymbol{\Sigma}}^2.$$

In this section we prove the following lemma.

Lemma 86. *If for some $k < n$ the matrix $\boldsymbol{A}_k$ is PD, then*

$$V \le \frac{\mu_1(\boldsymbol{A}_k^{-1})^2 tr(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top)}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2} + \mu_1(\boldsymbol{A}_k^{-1})^2 tr(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top).$$

Note that the RHS of the inequality above is straightforward to estimate if one knows the spectrum of $\boldsymbol{A}_k$. Indeed, the matrices $\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top$ and $\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top$ have i.i.d. elements on their diagonals, so their traces concentrate around expectations:

$$\operatorname{tr}(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top) \approx kn \text{ and } \operatorname{tr}(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \approx n\sum_{i>k}\lambda_i^2.$$

When it comes to the matrix $\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}/n$, this is just a sample covariance matrix of n isotropic vectors in k-dimensional space. Since k is small compared to n, it concentrates around the identity. Thus,

$$\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right) \approx n.$$

These computations are done rigorously in the proof of Theorem 20.

First k components

It was shown in Section A.6 that the following identity holds (c.f. (A.7)):

$$\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{\varepsilon} = \hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k} + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}.$$

Multiplying the identity by $\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}^\top$ from the left, and using that $\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}^\top\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k} \ge 0$ we get

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}^\top\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{\varepsilon} \ge \hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}^\top\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}. \tag{A.8}$$

The leftmost expression is linear in $\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}$, and the rightmost is quadratic. We use these expressions to bound $\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}$.

First, we extract that norm from the quadratic part

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}^\top\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k} \ge& \mu_n(\boldsymbol{A}_k^{-1})\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}^\top\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}\\ \ge& \mu_n(\boldsymbol{A}_k^{-1})\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right).\end{aligned}$$

Then we can substitute (A.8) and apply Cauchy-Schwarz to obtain

$$\begin{aligned}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}}\mu_n(\boldsymbol{A}_k^{-1})\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right) &\le \hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}^\top\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}\\ &\le \hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}^\top\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{\varepsilon}\\ &\le \|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}\left\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{\varepsilon}\right\|,\end{aligned}$$

and so

$$\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}} \le \frac{\boldsymbol{\varepsilon}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{\varepsilon}}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2}.$$

Since $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$, taking expectation in $\boldsymbol{\varepsilon}$ only leaves the trace in the numerator:

$$\begin{aligned}\frac{1}{v_{\boldsymbol{\varepsilon}}^2}\mathbb{E}_{\boldsymbol{\varepsilon}}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}} \le&\frac{\operatorname{tr}(\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2}\\ \le&\frac{\mu_1(\boldsymbol{A}_k^{-1})^2\operatorname{tr}(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top)}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2},\end{aligned}$$

where we transitioned to the second line by using the fact that $\operatorname{tr}(\boldsymbol{M}\boldsymbol{M}'\boldsymbol{M}) \le \mu_1(\boldsymbol{M})^2\operatorname{tr}(\boldsymbol{M}')$ for PD matrices $\boldsymbol{M}, \boldsymbol{M}'$.

Components starting from $k+1$-st

The rest of the variance term is

$$\left\|\boldsymbol{\Sigma}_{k:\infty}^{1/2}\boldsymbol{X}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}\right\|^2 = \boldsymbol{\varepsilon}^\top\boldsymbol{A}^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}.$$

Since $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$, taking expectation in $\boldsymbol{\varepsilon}$ only leaves the trace of the matrix:

$$\begin{aligned}\frac{1}{v_{\boldsymbol{\varepsilon}}^2}\mathbb{E}_{\boldsymbol{\varepsilon}}\left\|\boldsymbol{\Sigma}_{k:\infty}^{1/2}\boldsymbol{X}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}\right\|^2 =&\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\boldsymbol{A}^{-1})\\ \le&\mu_1(\boldsymbol{A}^{-1})^2\operatorname{tr}(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top)\\ \le&\mu_1(\boldsymbol{A}_k^{-1})^2\operatorname{tr}(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top).\end{aligned}$$

Here we again used the fact that $\operatorname{tr}(\boldsymbol{M}\boldsymbol{M}'\boldsymbol{M}) \le \mu_1(\boldsymbol{M})^2\operatorname{tr}(\boldsymbol{M}')$ for PD matrices $\boldsymbol{M}, \boldsymbol{M}'$ to transition to the second line. We then used $\boldsymbol{A} \succeq \boldsymbol{A}_k$ to infer $\mu_1(\boldsymbol{A}^{-1}) \le \mu_1(\boldsymbol{A}_k^{-1})$.

A.8 Bias

The bias term is given by $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2$. In this section we prove the following

Lemma 87 (Bias term)**.** *Suppose that for some $k < n$ the matrix $\boldsymbol{A}_k$ is PD. Then there exists an absolute constant c such that*

$$
\begin{aligned}
&\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*)\|_{\boldsymbol{\Sigma}}^2/c \\
&\le \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)}{\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2}\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2 \\
&\quad + \frac{\|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2} \\
&\quad + \lambda_{k+1}\big(1 + \max(0, -\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\mu_1(\boldsymbol{A}^{-1})\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2 \\
&\quad + \lambda_{k+1}\big(1 + \max(0, -\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})}{\mu_k(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})^2}\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}\|^2.
\end{aligned}
$$

First k components

We need to bound $\|\boldsymbol{\theta}^*_{0:k} - \hat{\boldsymbol{\theta}}(\boldsymbol{y})_{0:k}(\lambda, \boldsymbol{X}\boldsymbol{\theta}^*)\|^2_{\boldsymbol{\Sigma}_{0:k}}$. By Section A.6, in particular identity (A.7), we have

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*)_{0:k} + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*)_{0:k} = \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}\boldsymbol{\theta}^*.$$

Denote the error vector as $\boldsymbol{\zeta} := \hat{\boldsymbol{\theta}}(\boldsymbol{X}\boldsymbol{\theta}^*) - \boldsymbol{\theta}^*$. We can rewrite the equation above as

$$\boldsymbol{\zeta}_{0:k} + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\zeta}_{0:k} = \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty} - \boldsymbol{\theta}^*_{0:k}.$$

Multiplying both sides by $\boldsymbol{\zeta}_{0:k}^\top$ from the left and using that $\boldsymbol{\zeta}_{0:k}^\top\boldsymbol{\zeta}_{0:k} = \|\boldsymbol{\zeta}_{0:k}\|^2 \ge 0$ we obtain

$$\boldsymbol{\zeta}_{0:k}^\top\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\zeta}_{0:k} \le \boldsymbol{\zeta}_{0:k}^\top\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty} - \boldsymbol{\zeta}_{0:k}^\top\boldsymbol{\theta}^*_{0:k}.$$

Next, divide and multiply by $\boldsymbol{\Sigma}_{0:k}^{1/2}$ in several places:

$$
\begin{aligned}
\boldsymbol{\zeta}_{0:k}^\top\boldsymbol{\Sigma}_{0:k}^{1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\Sigma}_{0:k}^{1/2}\boldsymbol{\zeta}_{0:k} \le&\boldsymbol{\zeta}_{0:k}^\top\boldsymbol{\Sigma}_{0:k}^{1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty} \\
&- \boldsymbol{\zeta}_{0:k}^\top\boldsymbol{\Sigma}_{0:k}^{1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}.
\end{aligned}
$$

Now we pull out the lowest singular values of the matrices in the LHS and largest singular values of the matrices in the RHS to obtain lower and upper bounds respectively, yielding

$$
\begin{aligned}
&\|\boldsymbol{\zeta}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}}\mu_n(\boldsymbol{A}_k^{-1})\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right) \\
&\le \|\boldsymbol{\zeta}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}\mu_1(\boldsymbol{A}_k^{-1})\sqrt{\mu_1\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)}\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\| \\
&+ \|\boldsymbol{\zeta}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}\|\boldsymbol{\theta}^*_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}},
\end{aligned}
$$

and so

$$\begin{aligned}\|\boldsymbol{\zeta}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}} \le & \frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})} \frac{\mu_1\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top \boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^{1/2}}{\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top \boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)} \|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\| \\ & + \frac{\|\boldsymbol{\theta}^*_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}}{\mu_n(\boldsymbol{A}_k^{-1})\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top \boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)}.\end{aligned}$$

The rest of the components

Recall that the full bias term is $\|(\boldsymbol{I}_p - \boldsymbol{X}^\top(\lambda \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X})\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2$ and that $\boldsymbol{A} = \lambda\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top$. The contribution of the components of $\boldsymbol{\zeta}$, starting from the $k+1$st can be bounded as follows:

$$\begin{aligned}&\|\boldsymbol{\theta}^*_{k:\infty} - \boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \\ &\qquad \le 3\left(\|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\right).\end{aligned}$$

First of all, let's deal with the second term:

$$\begin{aligned}\|\boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} =& \|\boldsymbol{\Sigma}_{k:\infty}^{1/2}\boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2 \\ \le& \|\boldsymbol{\Sigma}_{k:\infty}\|\|\boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2 \\ =& \lambda_{k+1}(\boldsymbol{\theta}^*_{k:\infty})^\top \boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}^{-1}\underbrace{(\boldsymbol{A} - \lambda\boldsymbol{I}_n - \boldsymbol{X}_{0:k}\boldsymbol{X}_{0:k}^\top)}_{\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top}\boldsymbol{A}^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty} \\ \le& \lambda_{k+1}(\boldsymbol{\theta}^*_{k:\infty})^\top \boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}^{-1}\big(\boldsymbol{A} + \max(0, -\lambda)\boldsymbol{I}_n\big)\boldsymbol{A}^{-1}\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty} \\ \le& \lambda_{k+1}\big(\mu_1(\boldsymbol{A}^{-1}) + \max(0, -\lambda)\mu_1(\boldsymbol{A}^{-1})^2\big)\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2 \\ \le& \lambda_{k+1}\big(1 + \max(0, -\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\mu_1(\boldsymbol{A}_k^{-1})\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2,\end{aligned}$$

where we used that $\mu_1(\boldsymbol{A}_k^{-1}) \ge \mu_1(\boldsymbol{A}^{-1})$ in the last transition.

Now, let's deal with the last term. Note that $\boldsymbol{A} = \boldsymbol{A}_k + \boldsymbol{X}_{0:k}\boldsymbol{X}_{0:k}^\top$. By the Sherman–Morrison–Woodbury formula,

$$\begin{aligned}\boldsymbol{A}^{-1}\boldsymbol{X}_{0:k} =& (\boldsymbol{A}_k^{-1} + \boldsymbol{X}_{0:k}\boldsymbol{X}_{0:k}^\top)^{-1}\boldsymbol{X}_{0:k} \\ =& \left(\boldsymbol{A}_k^{-1} - \boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\left(I_k + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\right)^{-1}\boldsymbol{X}_{0:k}^T\boldsymbol{A}_k^{-1}\right)\boldsymbol{X}_{0:k} \\ =& \boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\left(\boldsymbol{I}_n - \left(I_k + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\right)^{-1}\boldsymbol{X}_{0:k}^T\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\right) \\ =& \boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\left(\boldsymbol{I}_n - \left(I_k + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\right)^{-1}\left(I_k + \boldsymbol{X}_{0:k}^T\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k} - I_k\right)\right) \\ =& \boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\left(I_k + \boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\right)^{-1}.\end{aligned}$$

Thus,

$$
\begin{aligned}
&\|\boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\
=&\|\boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\left(I_k+\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\right)^{-1}\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\
=&\|\boldsymbol{\Sigma}_{k:\infty}^{1/2}\boldsymbol{X}_{k:\infty}^\top \boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}\|^2\\
\le&\|\boldsymbol{A}_k^{-1/2}\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\boldsymbol{A}_k^{-1/2}\|\mu_1(\boldsymbol{A}_k^{-1/2})^2\frac{\mu_1(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})}{\mu_k(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})^2}\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}\|^2\\
\le&\|\boldsymbol{\Sigma}_{k:\infty}\|\|\boldsymbol{A}_k^{-1/2}\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\boldsymbol{A}_k^{-1/2}\|\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})}{\mu_k(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})^2}\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}\|^2\\
=&\lambda_1\|\boldsymbol{I}_n-\lambda\boldsymbol{A}_k^{-1}\|\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})}{\mu_k(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})^2}\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}\|^2\\
\le&\lambda_1\big(1+\max(0,-\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})}{\mu_k(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})^2}\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}\|^2,
\end{aligned}
$$

where in the last transition we used the fact that $\boldsymbol{I}_n-\lambda\boldsymbol{A}_k^{-1}$ is a PSD matrix with norm bounded by 1 for $\lambda>0$.

Putting those bounds together yields the result.

A.9 Main results

Upper bound on the prediction MSE

Theorem 20. *There exists a (large) constant c, which only depends on σ_x, s.t. for any $k<n/c$ with probability at least $1-ce^{-n/c}$, if the matrix $\boldsymbol{A}_k$ is PD, then*

$$
\begin{aligned}
B/c\le&\|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\left(1+\frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}+n\lambda_{k+1}\mu_1(\boldsymbol{A}_k^{-1})\left(1+\max(0,-\lambda)\mu_1(\boldsymbol{A}_k^{-1})\right)\right)\\
&+\|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\left(\frac{1}{n^2\mu_n(\boldsymbol{A}_k^{-1})^2}+\frac{\lambda_{k+1}}{n}\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\left(1+\max(0,-\lambda)\mu_1(\boldsymbol{A}_k^{-1})\right)\right),\\
V/c\le&\frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{k}{n}+n\mu_1(\boldsymbol{A}_k^{-1})^2\sum_{i>k}\lambda_i^2.
\end{aligned}
$$

Proof. Lemmas 86 and 87 bound the bias and variance on the event that $\boldsymbol{A}_k$ is PD. Next to those lemmas we already put explanations of why those bounds are easy to assess via concentration arguments. Here we just do this rigorously.

Recall the bounds from Lemmas 86 and 87: for some absolute constant c

$$B/c \leq \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \tag{A.9}$$

$$+\frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)}{\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2}\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2 \tag{A.10}$$

$$+\frac{\|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2} \tag{A.11}$$

$$+\lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\mu_1(\boldsymbol{A}^{-1})\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2 \tag{A.12}$$

$$+\lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})}{\mu_k(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})^2}\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}\|^2, \tag{A.13}$$

$$V/c \leq \frac{\mu_1(\boldsymbol{A}_k^{-1})^2\mathrm{tr}(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top)}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2} \tag{A.14}$$

$$+\mu_1(\boldsymbol{A}_k^{-1})^2\mathrm{tr}(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top), \tag{A.15}$$

where the first four terms correspond to the bias and the last two to the variance. By inspecting that expression one can notice that it consists of some products of simple quantities that could be assessed individually. Namely, those quantities are:

1. $\mu_1(\boldsymbol{A}_k^{-1})$ and $\mu_n(\boldsymbol{A}_k^{-1})$ — smallest and largest singular values of $\boldsymbol{A}_k$. In this theorem we assume that those quantities are known or there is some oracle control over them.

2. $\mu_1\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)$ and $\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)$.

   The matrix $\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2} \in \mathbb{R}^{k\times n}$ has n i.i.d. columns with isotropic sub-Gaussian distribution in $\mathbb{R}^k$. The matrix $\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}/n$ is the sample covariance matrix of those columns, so when $k \ll n$ it concentrates around its expectation, which is I_k. More precisely, by Theorem 5.39 in [56], for some constants c'_x, C'_x (which only depend on σ_x) for every $t > 0$ s.t. $\sqrt{n} - C'_x\sqrt{k} - \sqrt{t} > 0$, with probability $1 - 2\exp(-c'_x t)$,

$$\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right) \geq \left(\sqrt{n} - C'_x\sqrt{k} - \sqrt{t}\right)^2, \tag{A.16}$$

$$\mu_1\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right) \leq \left(\sqrt{n} + C'_x\sqrt{k} + \sqrt{t}\right)^2. \tag{A.17}$$

3. $\mathrm{tr}\left(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top\right)$ and $\mathrm{tr}\left(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\right)$.

   $\mathrm{tr}\left(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top\right)$ is the sum of squared norms of columns of $\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top$, which are n i.i.d. isotropic vectors in $\mathbb{R}^k$. Analogously, $\mathrm{tr}\left(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\right)$ is the sum of squared norms of n i.i.d. sub-Gaussian vectors with covariance $\boldsymbol{\Sigma}_{k:\infty}^2$. Therefore, they concentrate around their expectations by the law of large numbers. More precisely, by Lemma 81 with probability at least $1 - 4e^{-c_2 t}$,

$$\begin{aligned} \operatorname{tr}\left(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top\right) &\le (n + \sqrt{tn}\sigma_x^2)k, \\ \operatorname{tr}\left(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\right) &\le (n + \sqrt{tn}\sigma_x^2)\sum_{i>k}\lambda_i^2. \end{aligned}$$

4. $\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}_{k:\infty}^*\|^2$.

   Once again, this quantity concentrates by the law of large numbers. The vector $\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}_{k:\infty}^*/\|\boldsymbol{\theta}_{k:\infty}^*\|_{\boldsymbol{\Sigma}_{k:\infty}}$ has n i.i.d. centered components with unit variances and sub-Gaussian norms at most σ_x. Treating those components as sub-Gaussian vectors in $\mathbb{R}^1$, we can apply Lemma 81 to get that for any $t \in (0, n)$, with probability at least $1 - 2e^{-c_2 t}$,

$$\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}_{k:\infty}^*\|^2 \le (n + \sqrt{tn}\sigma_x^2)\|\boldsymbol{\theta}_{k:\infty}^*\|_{\boldsymbol{\Sigma}_{k:\infty}}^2.$$

Now take constant c_4 to be large enough depending on σ_x and set $t = n/c_4$. For some constant c_5 which only depends on σ_x we get that with probability at least $1 - c_5 e^{-n/c_5}$, all the following inequalities hold at the same time:

$$\begin{aligned} \mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right) &\ge n/c_5, \\ \mu_1\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right) &\le c_5 n, \\ \|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}_{k:\infty}^*\|^2 &\le c_5 n\|\boldsymbol{\theta}_{k:\infty}^*\|_{\boldsymbol{\Sigma}_{k:\infty}}^2, \\ \|\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}&, \\ \operatorname{tr}\left(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top\right) &\le c_5 nk, \\ \operatorname{tr}\left(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top\right) &\le c_5 n\sum_{i>k}\lambda_i^2. \end{aligned} \tag{A.18}$$

Next, plug these bounds into (A.10)–(A.15):

$$\frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)}{\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2}\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}_{k:\infty}^*\|^2 \le c_5^3\frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}\|\boldsymbol{\theta}_{k:\infty}^*\|_{\boldsymbol{\Sigma}_{k:\infty}}^2,$$

$$\frac{\|\boldsymbol{\theta}_{0:k}^*\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2} \le c_5^2\frac{\|\boldsymbol{\theta}_{0:k}^*\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2}{\mu_n(\boldsymbol{A}_k^{-1})^2 n^2},$$

$$\begin{aligned} &\lambda_{k+1}\big(1 + \max(0, -\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\mu_1(\boldsymbol{A}^{-1})\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}_{k:\infty}^*\|^2 \le \\ &\qquad\qquad \le c_5^2\lambda_{k+1}\left(1 + \max(0, -\lambda)\mu_1(\boldsymbol{A}_k^{-1})\right)\mu_1(\boldsymbol{A}_k^{-1})n\|\boldsymbol{\theta}_{k:\infty}^*\|_{\boldsymbol{\Sigma}_{k:\infty}}^2, \end{aligned}$$

$$\begin{aligned}
\lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(\boldsymbol{A}_k^{-1})\big)\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})}{\mu_k(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2})^2}\|\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\theta}^*_{0:k}\|^2 \le&\\
\le c_5^4\lambda_{k+1}\left(1+\max(0,-\lambda)\mu_1(\boldsymbol{A}_k^{-1})\right)\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{1}{n}\|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}&,
\end{aligned}$$

$$\frac{\mu_1(\boldsymbol{A}_k^{-1})^2\mathrm{tr}(\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1}\boldsymbol{X}_{0:k}^\top)}{\mu_n(\boldsymbol{A}_k^{-1})^2\mu_k\left(\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{X}_{0:k}^\top\boldsymbol{X}_{0:k}\boldsymbol{\Sigma}_{0:k}^{-1/2}\right)^2} \le c_5^3\frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{k}{n},$$

$$\mu_1(\boldsymbol{A}_k^{-1})^2\mathrm{tr}(\boldsymbol{X}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \le c_5\mu_1(\boldsymbol{A}_k^{-1})^2 n\sum_{i>k}\lambda_i^2.$$

Putting all the terms together gives the result. □

Corollary 21. *Fix any constants $\gamma \in [0,1)$ and $L > 0$. There exists a constant c that only depends on σ_x, γ, L s.t. for any $k < n/c$ and $\delta < 1 - ce^{-n/c}$ under assumptions NoncritReg(k,γ) and CondNum(k,δ,L), it holds that $\rho_k > c^{-1}$, and with probability at least $1-\delta-ce^{-n/c}$,*

$$\begin{aligned}
B/c \le&\|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\left(\frac{\lambda+\sum_{i>k}\lambda_i}{n}\right)^2,\\
V/c \le&\frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda+\sum_{i>k}\lambda_i\right)^2}.
\end{aligned}$$

Proof. Almost all the work was already done in Lemma 85. It says that for some absolute constant c_1 and for any $t \in (0,n)$ with probability at least $1-\delta-2e^{-c_1 t}$,

$$\begin{aligned}
\mu_n(\boldsymbol{A}_k) \ge&\frac{1}{L}\left(1-\frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda+\sum_i\lambda_i\right),\\
\mu_1(\boldsymbol{A}_k) \le&L\left(1-\frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda+\sum_i\lambda_i\right).
\end{aligned}$$

Moreover, if $\delta < 1-4e^{-c_1 t}$, then

$$\rho_k \ge \frac{1-\sigma^2\sqrt{t/n}}{L+\frac{\gamma}{1-\gamma}+\frac{\sqrt{t}\sigma^2 L}{\sqrt{n}(1-\gamma)}}.$$

We just need to choose t, plug these bounds into the result of Theorem 20 and evaluate the result up to multiplicative constants.

First, choose constant c_2 large enough depending on L, γ, σ_x , and put $t = n/c_2$. Statements above imply that if $\delta < 1 - 4e^{-n/(c_1c_2)}$, then for some constant c_3 which only depends on L, γ, σ_x, with probability at least $1 - \delta - c_2e^{-n/(c_1c_2)}$,

$$\begin{aligned}\mu_n(\boldsymbol{A}_k^{-1}) =&\mu_1(\boldsymbol{A}_k)^{-1} \geq \frac{1}{c_3}\left(\lambda + \sum_i \lambda_i\right)^{-1},\\ \mu_1(\boldsymbol{A}_k^{-1}) =&\mu_n(\boldsymbol{A}_k)^{-1} \leq \frac{1}{c_3}\left(\lambda + \sum_i \lambda_i\right)^{-1},\\ \rho_k \geq&\frac{1}{c_3}.\end{aligned}$$

These three inequalities allow us to evaluate the result of Theorem 20: let's plug them term-by-term:

- Since $\lambda > -\gamma\sum_{i>k}\lambda_i$,

  $$\max(0, -\lambda) \leq \frac{\gamma}{1-\gamma}\left(\lambda + \sum_i \lambda_i\right).$$

  Thus,

  $$1 + \max(0, -\lambda)\mu_1(\boldsymbol{A}_k^{-1}) \leq 1 + \frac{\gamma}{1-\gamma}c_3,$$

  so this term is just a constant.

- $$n\lambda_{k+1}\mu_1(\boldsymbol{A}_k^{-1}) \leq c_3n\lambda_{k+1}\left(\lambda + \sum_i \lambda_i\right) = c_3/\rho_k \leq c_3^2,$$

  so this term is also just a constant.

- $$\frac{1}{n^2\mu_n(\boldsymbol{A}_k^{-1})^2} \leq \frac{c_3^2}{n}\left(\lambda + \sum_i \lambda_i\right)^2.$$

- $$\begin{aligned}\frac{\lambda_{k+1}}{n}\frac{\mu_1(\boldsymbol{A}_k^{-1})}{\mu_n(\boldsymbol{A}_k^{-1})^2} \leq&\frac{c_3^3}{n^2}\cdot n\lambda_{k+1}\left(\lambda + \sum_i \lambda_i\right)\\ =&\frac{c_3^3}{n^2}\cdot\rho_k^{-1}\left(\lambda + \sum_i \lambda_i\right)^2\\ \leq&\frac{c_3^4}{n^2}\left(\lambda + \sum_i \lambda_i\right)^2.\end{aligned}$$

- $\frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2} \leq L^2$ — also just a constant.
- $$n\mu_1(\boldsymbol{A}_k^{-1})^2 \leq c_3^2 n \left(\lambda + \sum_i \lambda_i\right)^{-2}.$$

Plugging all these bounds in the statement of Theorem 20 gives the result for a large enough c. □

Upper bound matches the lower bound

In the next theorem we show that the upper bound given in Theorem 20 matches the lower bounds from Lemmas 22 and 23 if we choose suitable k. Note that by Lemmas 85 and 26, being able to control the condition number of $\boldsymbol{A}_{k'}$ for some $k' < n$ implies that we can choose a suitable k.

Theorem 25 (The lower bound is the same as the upper bound)**.** *Denote*

$$\begin{aligned}
\underline{B} &:= \sum_i \frac{\lambda_i |\theta_i^*|^2}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2},\\
\overline{B} &:= \|\boldsymbol{\theta}_{k:\infty}^*\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 + \|\boldsymbol{\theta}_{0:k}^*\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2 \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2,\\
\underline{V} &:= \frac{1}{n}\sum_i \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\},\\
\overline{V} &:= \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}.
\end{aligned}$$

Fix constants $a > 0$ and $b > 1/n$. There exists a constant $c > 0$ that only depends on a, b, s.t. the following holds: if either $\rho_k \in (a, b)$ or $k = \min\{\kappa : \rho_\kappa > b\}$, then

$$c^{-1} \leq \underline{B} \,/\, \overline{B} \leq 1, \quad c^{-1} \leq \underline{V} \,/\, \overline{V} \leq 1.$$

Proof. First of all, we represent

$$\begin{aligned}
\|\boldsymbol{\theta}_{k:\infty}^*\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 + \|\boldsymbol{\theta}_{0:k}^*\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2 \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2 &= \sum_i \left(\mathbb{1}\{i \leq k\}\frac{|\theta_i^*|^2 \rho_k^2 \lambda_{k+1}^2}{\lambda_i} + \mathbb{1}\{i > k\}\lambda_i |\theta_i^*|^2\right)\\
\frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2} &= \sum_i \left(\mathbb{1}\{i \leq k\}\frac{1}{n} + \mathbb{1}\{i > k\}\frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2}\right)
\end{aligned}$$

In the following we will bound the ratio of the sums from the statement of the theorem by bounding the ratios of the corresponding terms.

- First case: $\rho_k \in (a, b)$.
  - Bias term:
    - $i \le k$:

$$\begin{aligned}
&\frac{\lambda_i|\theta_i^*|^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2} : \frac{|\theta_i^*|^2\rho_k^2\lambda_{k+1}^2}{\lambda_i}\\
=&\frac{\lambda_i^2}{\rho_k^2\lambda_{k+1}^2\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2}\\
=&\left(1+\frac{\lambda_{k+1}\rho_k}{\lambda_i}\right)^{-2}\\
\in&\left((1+b)^{-2}, 1\right)
\end{aligned}$$

    - $i > k$:

$$\begin{aligned}
&\frac{\lambda_i|\theta_i^*|^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2} : \lambda_i|\theta_i^*|^2\\
=&\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^{-2}\\
\in&\left((1+a^{-1})^{-2}, 1\right)
\end{aligned}$$

  - Variance term:
    - $i \le k$:

$$\begin{aligned}
&\frac{1}{n}\min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\} : \frac{1}{n}\\
\in&\left((1+b)^{-2}, 1\right]
\end{aligned}$$

    - $i > k$:

$$\begin{aligned}
&\frac{1}{n}\min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\} : \frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2}\\
=&\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2} : \frac{\lambda_i^2}{\lambda_{k+1}^2\rho_k^2}\\
=&\frac{\rho_k^2}{(\rho_k+1)^2}\\
\in&\left((1+a^{-1})^{-2}, 1\right)
\end{aligned}$$

- Second case: $k = \min\{l : \rho_l > b\}$. In this case we have

$$\begin{gathered}\rho_k \geq b,\\ \frac{\lambda_k + n\lambda_{k+1}\rho_k}{n\lambda_k} = \frac{\lambda + \lambda_k + \sum_{i>k}\lambda_i}{n\lambda_k} = \rho_{k-1} < b,\\ \forall i \leq k: \quad \lambda_i \geq \lambda_k \geq \frac{n\lambda_{k+1}\rho_k}{nb-1} = \frac{\lambda_{k+1}\rho_k}{b} \geq \frac{\lambda_{k+1}\rho_k}{b}.\end{gathered}$$

The rest of the computation is analogous to the previous case:

- Bias term:
  - $i \leq k$:

$$\begin{aligned}&\frac{\lambda_i|\theta_i^*|^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2} : \frac{|\theta_i^*|^2\rho_k^2\lambda_{k+1}^2}{\lambda_i}\\ =&\frac{\lambda_i^2}{\rho_k^2\lambda_{k+1}^2\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2}\\ =&\left(1+\frac{\lambda_{k+1}\rho_k}{\lambda_i}\right)^{-2}\\ \in&\left[(1+b)^{-2}, 1\right)\end{aligned}$$

  - $i > k$:

$$\begin{aligned}&\frac{\lambda_i|\theta_i^*|^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2} : \lambda_i|\theta_i^*|^2\\ =&\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^{-2}\\ \in&\left[(1+b^{-1})^{-2}, 1\right)\end{aligned}$$

- Variance term:
  - $i \leq k$:

$$\begin{aligned}&\frac{1}{n}\min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\} : \frac{1}{n}\\ \in&\left[\frac{\lambda_{k+1}^2\rho_k^2/b^2}{\lambda_{k+1}^2(\rho_k+1)^2}, 1\right]\\ \subseteq&\left[\frac{b^2}{(b+1)^2b^2}, 1\right]\\ =&\left[(b+1)^{-2}, 1\right]\end{aligned}$$

$*$ $i > k$:

$$\begin{aligned}&\frac{1}{n}\min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\} : \frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2}\\ =&\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2} : \frac{\lambda_i^2}{\lambda_{k+1}^2\rho_k^2}\\ =&\frac{\rho_k^2}{(\rho_k+1)^2}\\ \in&\left[(1+b^{-1})^{-2}, 1\right]\end{aligned}$$

□

Alternative form of the main bound

Lemma 27. *Suppose $k < n/c$ for some $c > 1$ and $k^* < k$. Then*

$$\lambda_{k+1}\rho_k \leq \lambda_{k^*+1}\rho_{k*} \leq \lambda_{k+1}\rho_k/(1 - b^{-1}c^{-1}).$$

Proof.

$$\begin{aligned}\lambda_{k^*+1}\rho_{k*} =&\lambda_{k+1}\rho_k + \frac{1}{n}\sum_{i=k^*+1}^{k}\lambda_i\\ \leq&\lambda_{k+1}\rho_k + \frac{k-k^*}{n}\lambda_{k^*+1}\\ =&\lambda_{k+1}\rho_k + \frac{k-k^*}{n}\frac{\lambda_{k^*+1}\rho_{k*}}{\rho_{k*}}\\ \leq&\lambda_{k+1}\rho_k + \frac{\lambda_{k^*+1}\rho_{k*}}{bc},\end{aligned}$$

where we used $k - k^* < n/c$ and $\rho_{k^*} > b$ in the last transition. Moving $\frac{\lambda_{k^*+1}\rho_{k*}}{bc}$ to the left-hand side and dividing both sides by $(1 - b^{-1}c^{-1})$ gives the result. □

Corollary 28. *There is a large positive constant c that only depends on σ_x such that if*

$$\lambda > cn\lambda_{\lfloor n/c\rfloor} + 2\sum_{i>\lfloor n/c\rfloor}\lambda_i,$$

then

$$\begin{aligned}B/c \leq&\sum_i \lambda_i|\theta_i^*|^2\frac{(\lambda/n)^2}{\left(\lambda/n+\lambda_i\right)^2},\\ V/c \leq&\frac{1}{n}\sum_i\frac{\lambda_i^2}{\left(\lambda/n+\lambda_i\right)^2}.\end{aligned}$$

Proof. Set $\gamma = 0$ and denote c_1 to be the constant c from Lemma 18. Take $L = 2c_1$ and $b = L^2$, $a = b/2$. For such choice of γ, L, a, b denote c_2 to be the constant from Theorem 16 and take any $\tilde{k} < n/c_2$.

Take any λ s.t.

$$\lambda \geq 2\sum_{i>\tilde{k}} \lambda_i \quad \text{and} \quad \rho_{\tilde{k}} \geq L^2,$$

i.e.,

$$\lambda \geq \max\left(2\sum_{i>\tilde{k}} \lambda_i,\ L^2 n\lambda_{\tilde{k}+1} - \sum_{i>\tilde{k}} \lambda_i\right).$$

Then the conditions of the first part of Lemma 18 are satisfied with $\delta = 0$, which means that with probability $1 - c_1 e^{-n/c_1}$, $\mu_n(\boldsymbol{A}_{\tilde{k}}) \geq L^{-1}\mu_1(\boldsymbol{A}_{\tilde{k}})$, so the assumptions of the first part of Theorem 16 are satisfied with $\delta = c_1 e^{-n/c_1}$ and $k = \tilde{k}$. Note also that since $\rho_{\tilde{k}} \geq L^2 = b$, then $k^* \leq \tilde{k}$. This means that with probability at least $1 - c_1 e^{-n/c_1} - c_2 e^{-n/c_2}$, for $k = k^*$,

$$\begin{aligned}
B/c_2 \leq& \|\boldsymbol{\theta}^*_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{\theta}^*_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}}\left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2,\\
V/c_2 \leq& \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}.
\end{aligned}$$

Now since $k = k^*$, by Theorem 25 there exists a large constant c_3 (that depends on b and c_2) such that on the same event,

$$\begin{aligned}
B/c_3 \leq& \sum_i \lambda_i |\theta^*_i|^2 \frac{\rho_k^2 \lambda_{k^*+1}^2}{\left(\rho_{k^*}\lambda_{k^*+1} + \lambda_i\right)^2},\\
V/c_3 \leq& \frac{1}{n}\sum_i \frac{\lambda_i^2}{\left(\rho_{k^*}\lambda_{k^*+1} + \lambda_i\right)^2},
\end{aligned}$$

where $\tilde{B}$ and $\tilde{V}$ are defined in Equations (2.9)–(2.10).

We've just cast the bounds to the alternative form, which allows us to transition from k^* to the initial value $\tilde{k}$. By Lemma 27 since $n/c_2 \geq \tilde{k} \geq k^*$ there exists a constant c_4 that depends on c_2, c_3, b such that on the same event

$$\begin{aligned}
B/c_4 \leq& \sum_i \lambda_i |\theta^*_i|^2 \frac{\rho_k^2 \lambda_{k+1}^2}{\left(\rho_k\lambda_{k+1} + \lambda_i\right)^2},\\
V/c_4 \leq& \frac{1}{n}\sum_i \frac{\lambda_i^2}{\left(\rho_k\lambda_{k+1} + \lambda_i\right)^2}.
\end{aligned}$$

Finally, since $\lambda > 2\sum_{i>k}\lambda_i$, we have

$$\lambda/n \le \rho_k \lambda_{k+1} = \frac{1}{n}\left(\lambda + \sum_{i>k}\lambda_i\right) \le 1.5\lambda/n.$$

Thus, on the same event

$$\begin{aligned} B/(2.25c_4) \le & \sum_i \lambda_i |\theta_i^*|^2 \frac{(\lambda/n)^2}{(\lambda/n + \lambda_i)^2}, \\ V/(2.25c_4) \le & \frac{1}{n}\sum_i \frac{\lambda_i^2}{(\lambda/n + \lambda_i)^2}. \end{aligned}$$

To finish the proof take $c = \max(2.25c_4, c_1 + c_2, L^2)$ and $\tilde{k} = \lfloor n/c \rfloor$.

□

Lemma 29. *Suppose that $n \ge c^2 + c$ for some $c > 0$ and take*

$$\lambda = cn\lambda_{\lfloor n/c \rfloor} + 2\sum_{i > \lfloor n/c \rfloor}\lambda_i.$$

Then

$$d(\lambda/n) \ge \frac{n}{2\max(2, (c+1)^2)}.$$

Proof.

$$d(\lambda/n) = \sum_i \frac{\lambda_i}{\lambda_i + c\lambda_{\lfloor n/c \rfloor} + \frac{2}{n}\sum_{i>\lfloor n/c \rfloor}\lambda_i}.$$

Consider two cases:

Case 1: $(1+c)\lambda_{\lfloor n/c \rfloor} \ge \frac{2}{n}\sum_{i>\lfloor n/c \rfloor}\lambda_i$. Then

$$\begin{aligned} & \sum_i \frac{\lambda_i}{\lambda_i + c\lambda_{\lfloor n/c \rfloor} + \frac{2}{n}\sum_{i>\lfloor n/c \rfloor}\lambda_i} \\ \ge & \sum_i \frac{\lambda_i}{\lambda_i + (1+2c)\lambda_{\lfloor n/c \rfloor}} \\ \ge & \sum_{i \le \lfloor n/c \rfloor} \frac{\lambda_i}{\lambda_i + (1+2c)\lambda_{\lfloor n/c \rfloor}} \\ \ge & \sum_{i \le \lfloor n/c \rfloor} \frac{\lambda_i}{\lambda_i(2+2c)} \\ = & \frac{\lfloor n/c \rfloor}{2+2c}. \end{aligned}$$

Case 2: $(1+c)\lambda_{\lfloor n/c\rfloor} < \frac{2}{n}\sum_{i>\lfloor n/c\rfloor}\lambda_i$. Then

$$\begin{aligned}&\sum_i \frac{\lambda_i}{\lambda_i + c\lambda_{\lfloor n/c\rfloor} + \frac{2}{n}\sum_{i>\lfloor n/c\rfloor}\lambda_i}\\ \geq &\sum_{i>\lfloor n/c\rfloor} \frac{\lambda_i}{\lambda_i + c\lambda_{\lfloor n/c\rfloor} + \frac{2}{n}\sum_{i>\lfloor n/c\rfloor}\lambda_i}\\ \geq &\sum_{i>\lfloor n/c\rfloor} \frac{\lambda_i}{(1+c)\lambda_{\lfloor n/c\rfloor} + \frac{2}{n}\sum_{i>\lfloor n/c\rfloor}\lambda_i}\\ \geq &\sum_{i>\lfloor n/c\rfloor} \frac{\lambda_i}{\frac{4}{n}\sum_{i>\lfloor n/c\rfloor}\lambda_i}\\ = &\frac{n}{4}.\end{aligned}$$

A straightforward computation shows that if $n \geq c^2 + c$ then $n/c - 1 \geq n/(c+1)$, so

$$\frac{\lfloor n/c\rfloor}{2+2c} \geq \frac{n}{2(c+1)^2},$$

which finishes the proof.

□

A.10 Negative regularization

Lemma 32 (Lower bound on the bias for any non-negative regularization)**.** *There exist constants b, c that only depend on σ_x such that the following holds: suppose that assumptions IndepCoord and PriorSigns($\bar{\boldsymbol{\theta}}$) hold. Take $k = \min\{\kappa : \rho_\kappa(0) > b\}$ and suppose that $k > 0$. Then with probability at least $1 - ce^{-n/c}$ for any $\lambda \geq 0$*

$$\mathbb{E}_{\boldsymbol{\theta}^*} B \geq \frac{1}{c}\|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \frac{\left(\sum_{i>k}\lambda_i\right)^2}{n^2}.$$

Proof. We start exactly as in the proof of Lemma 23, where it was shown that if $\boldsymbol{A}_{-i}$ is PSD for every i (which is satisfied almost surely when $\lambda \geq 0$) then

$$\mathbb{E}_{\boldsymbol{\theta}^*} B \geq \sum_i \frac{\lambda_i \bar{\boldsymbol{\theta}}_i^2}{(1+\lambda_i \boldsymbol{z}_i^\top \boldsymbol{A}_{-i}^{-1}\boldsymbol{z}_i)^2} \geq \sum_i \frac{\lambda_i \bar{\boldsymbol{\theta}}_i^2}{(1+\lambda_i \mu_n(\boldsymbol{A}_{-i}^{-1})\|\boldsymbol{z}_i\|^2)^2}. \tag{A.19}$$

Note that have $\mu_n(\boldsymbol{A}_{-i}^{-1})$ is a decreasing function of λ with probability 1. Thus, the right-hand side of (A.19) is a non-decreasing function of λ with probability 1, and any lower bound for it when $\lambda = 0$ will also hold uniformly for all $\lambda \geq 0$. Thus, for the remainder of the proof, fix $\lambda = 0$.

We are going to use Lemma 31 to lower bound $\mu_n(\boldsymbol{A}_{-i})$ for each i separately (we are *not* looking for a uniform bound over all i simultaneously). If $i \le k$, then $\boldsymbol{A}_{-i} \succeq \boldsymbol{A}_k$ with probability 1, so we can just use Lemma 31 directly. If $i > k$, consider the following matrix:

$$\boldsymbol{X}_{k:\infty}^{(i)} := [\sqrt{\lambda_{k+1}}\boldsymbol{z}_{k+1}, \dots, \sqrt{\lambda_{i-1}}\boldsymbol{z}_{i-1}, \sqrt{\lambda_i}\boldsymbol{z}_1, \sqrt{\lambda_{i+1}}\boldsymbol{z}_{i+1}, \dots, \sqrt{\lambda_p}\boldsymbol{z}_p].$$

In words, we took matrix $\boldsymbol{X}$, multiplied the first column by $\sqrt{\lambda_i/\lambda_1}$ (to make the variances equal to λ_i), swapped the first column with the i-th column and dropped the first k columns. The purpose of this matrix is to write the following:

$$\boldsymbol{A}_{-i} = \sum_{j \ne i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top \succeq \lambda_i \boldsymbol{z}_1 \boldsymbol{z}_1^\top + \sum_{j>k, j\ne i} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^\top = \boldsymbol{X}_{k:\infty}^{(i)}(\boldsymbol{X}_{k:\infty}^{(i)})^\top.$$

Thus, to lower bound $\mu_n(\boldsymbol{A}_{-i})$ one can just lower bound $\mu_n(\boldsymbol{X}_{k:\infty}^{(i)}(\boldsymbol{X}_{k:\infty}^{(i)})^\top)$. This can be done by using Lemma 31 with matrix $\boldsymbol{X}_{k:\infty}^{(i)}$ instead of $\boldsymbol{X}_{k:\infty}$, which is valid because matrix $\boldsymbol{X}_{k:\infty}^{(i)}$ satisfies exactly the same assumptions, namely the matrix $\boldsymbol{X}_{k:\infty}^{(i)}\boldsymbol{\Sigma}_{k:\infty}^{-1/2}$ has independent centered σ_x-sub-Gaussian elements with unit variances.

Therefore, by Lemma 31 for some constant c_1 that only depends on σ_x for any i with probability at least $1 - c_1 e^{-n/c_1}$,

$$\begin{aligned}
\mu_n(\boldsymbol{A}_{-i}) \ge & \sum_{i>k} \lambda_i - c_1 \left(n\lambda_{k+1} + \sqrt{n \sum_{i>k} \lambda_i^2} \right) \\
\ge & \left(1 - c_1 \rho_k(0)^{-1} - c_1 \rho_k(0)^{-1/2}\right) \sum_{i>k} \lambda_i \\
= & n\lambda_{k+1}(\rho_k - c_1 - c_1\sqrt{\rho_k}),
\end{aligned}$$

where we used Equations (2.16) and (2.17). Choose a constant b large enough depending on c_1, so that $\rho_k - c_1 - c_1\sqrt{\rho_k} \ge \rho_k/c_2$ for some constant c_2 that only depends on σ_x.

By Lemma 81, for some absolute constant c_3 for any $t \in (0, n)$, w.p. at least $1 - 2e^{-t/c_3}$, we have $\|\boldsymbol{z}_i\|^2 \le n - \sqrt{tn}\sigma_x^2 \le n/2$, provided $t \le n/(4\sigma_x^4)$. Combining it with the previous results and taking constant c_4 large enough depending on σ_x and c_2 we get that if $\rho_k > c_4$ then for any i with probability at least $1 - c_4 e^{-n/c_4}$,

$$\frac{\lambda_i \bar{\boldsymbol{\theta}}_i^2}{(1 + \lambda_i \mu_n(\boldsymbol{A}_{-i}^{-1})\|\boldsymbol{z}_i\|^2)^2} \ge \frac{1}{c_4} \frac{\lambda_i \bar{\boldsymbol{\theta}}_i^2}{(1 + \frac{n\lambda_i}{n\lambda_{k+1}\rho_k})^2} = \frac{1}{c_4} \frac{\lambda_i \bar{\boldsymbol{\theta}}_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k})^2}.$$

Now we convert the high-probability lower bound for each term into the high-probability lower bound for the whole sum. Using Lemma 77 gives that with probability at least $1 - 2c_4 e^{-n/c_4}$,

$$\mathbb{E}_{\boldsymbol{\theta}^*} B \ge \frac{1}{2c_4} \sum_i \frac{\lambda_i \bar{\boldsymbol{\theta}}_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k})^2}.$$

Finally, by Theorem 25 there exists a constant c_5 that only depends on b s.t.

$$\sum_i \frac{\lambda_i \bar{\boldsymbol{\theta}}_i^2}{(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k})^2} \geq \frac{1}{c_5}\|\boldsymbol{\theta}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\left(\frac{\sum_{i>k}\lambda_i}{n}\right)^2.$$

Therefore, setting the constant c large enough (depending on b and σ_x) gives the result. □

Lemma 33 (Upper bound on excess risk for some negative regularization)**.** *There exists a constant c that only depends on σ_x such that the following holds: suppose that assumptions PriorSigns($\bar{\boldsymbol{\theta}}$) and IndepCoord hold and that $\rho_k(0) > c$ for some $k < n/c$. Assume also that*

$$v_{\boldsymbol{\varepsilon}}^2 \leq \frac{1}{c}\|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\frac{\left(\sum_{i>k}\lambda_i\right)^2}{n^3\left(\sum_{i>k}\lambda_i^2\right)^2}. \tag{2.18}$$

Then there exists such $\lambda < 0$ that with probability at least $1 - ce^{-n/c}$

$$\mathbb{E}_{\boldsymbol{\theta}^*}B + v_{\boldsymbol{\varepsilon}}^2 V \leq c\left(v_{\boldsymbol{\varepsilon}}^2\frac{k}{n} + v_{\boldsymbol{\varepsilon}}\|\bar{\boldsymbol{\theta}}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}\sqrt{\frac{\sum_{i>k}\lambda_i^2}{n}} + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n} + \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\right).$$

Proof. In the following $c_1, c_2, \dots$ are constants that only depend on σ_x.

Let's introduce a new variable $\diamondsuit$ such that $\lambda = -\sum_{i>k}\lambda_i + \diamondsuit$.

By Lemma 31 with probability at least $1 - c_1 e^{-n/c_1}$,

$$\mu_1(\boldsymbol{A}_k) = \lambda + \mu_1(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \leq \diamondsuit + c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right),$$

$$\mu_n(\boldsymbol{A}_k) = \lambda + \mu_n(\boldsymbol{X}_{k:\infty}\boldsymbol{X}_{k:\infty}^\top) \geq \diamondsuit - c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

Let's put

$$\sum_{i>k}\lambda_i > \diamondsuit > 2c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right). \tag{A.20}$$

Note that the range for $\diamondsuit$ is non-empty if ρ_k is large enough according to Equations (2.16) and (2.17). On the same event we get

$$\mu_n(\boldsymbol{A}_k^{-1})^{-1} = \mu_1(\boldsymbol{A}_k) \leq \frac{3}{2}\diamondsuit, \qquad \mu_n(\boldsymbol{A}_k^{-1}) \geq \frac{2}{3}\diamondsuit^{-1},$$

$$\mu_1(\boldsymbol{A}_k^{-1})^{-1} = \mu_n(\boldsymbol{A}_k) \geq \frac{1}{2}\diamondsuit, \qquad \mu_1(\boldsymbol{A}_k^{-1}) \leq 2\diamondsuit^{-1}.$$

Now we are in a position to use Theorem 20. Recall that $0 < \Diamond < \sum_{i>k}\lambda_i$. Thus

$$\max(0, -\lambda) = -\lambda = \sum_{i>k}\lambda_i - \Diamond \le \sum_{i>k}\lambda_i.$$

Note that results of Theorem 20 still apply for the case when the expectation of the bias term is taken over the prior from assumption *PriorSigns*($\bar{\boldsymbol{\theta}}$). Indeed, as explained in the sketch of its proof, it decomposes very clearly into an algebraic and a stochastic part, where concentration results are applied. One can see that the only stochastic quantity that changes when the expectation over $\boldsymbol{\theta}^*$ is taken is $\|\boldsymbol{X}_{0:k}\boldsymbol{\theta}^*_{0:k}\|^2$. To obtain the result of the theorem one needs to show that $\mathbb{E}^*_{\boldsymbol{\theta}}\|\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}\|^2 \le \tilde{c}\|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}$ with probability $1 - \tilde{c}e^{-n/\tilde{c}}$ for some $\tilde{c}$ that only depends on σ_x. This is indeed the case because expectations over $\boldsymbol{\theta}^*$ of the squared components of $\boldsymbol{X}_{k:\infty}\boldsymbol{\theta}^*_{k:\infty}$ are i.i.d. sub-Exponential random variables with expectation $\|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}$ and sub-Exponential norm bounded by $\bar{c}\|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}$ for a constant $\bar{c}$ that only depends on σ_x. Thus, the desired concentration result holds by the same application of Bernstein's inequality as in Lemma 81.

Thus, we can plug our bounds on eigenvalues into Theorem 20 to get that if $k < n/c_2$ then with probability at least $1 - c_1e^{-n/c_1} - c_2e^{-n/c_2}$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}B/c_2 \le & \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\left(1 + \frac{(2\Diamond^{-1})^2}{\left(\frac{2}{3}\Diamond^{-1}\right)^2} + n\lambda_{k+1}(2\Diamond^{-1})\left(1 + (2\Diamond^{-1})\sum_{i>k}\lambda_i\right)\right)\\
& + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}}\left(\frac{1}{n^2\left(\frac{2}{3}\Diamond^{-1}\right)^2} + \frac{\lambda_{k+1}}{n}\frac{(2\Diamond^{-1})}{\left(\frac{2}{3}\Diamond^{-1}\right)^2}\left(1 + (2\Diamond^{-1})\sum_{i>k}\lambda_i\right)\right),\\
V/c_2 \le & \frac{(2\Diamond^{-1})^2}{\left(\frac{2}{3}\Diamond^{-1}\right)^2}\frac{k}{n} + n(2\Diamond^{-1})^2\sum_{i>k}\lambda_i^2.
\end{aligned}$$

Recall that $\Diamond < \sum_{i>k}\lambda_i$, so $1 + (2\Diamond^{-1})\sum_{i>k}\lambda_i$ is the same as $\Diamond^{-1}\sum_{i>k}\lambda_i$ up to a constant multiplier. That is, on the same event,

$$\begin{aligned}
B/c_3 \le & \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\left(1 + \frac{n\lambda_{k+1}\sum_{i>k}\lambda_i}{\Diamond^2}\right) && \text{(A.21)}\\
& + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}}\left(\frac{\Diamond^2}{n^2} + \frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n}\right), && \text{(A.22)}\\
V/c_3 \le & \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\Diamond^2}. && \text{(A.23)}
\end{aligned}$$

One can see that $\Diamond$ balances the bias in the first k components against two things: the bias in the tail and the variance. The value of $\Diamond$ that is optimal to balance the bias in the first k components and the bias in the tail is $\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}$. As we will check further, up to a constant factor, $\Diamond$ will be in the range that we set in Equation (A.20). There are two cases then: the first case is when this choice of $\Diamond$ is optimal because the variance is not larger than the bias. The second case is when $\Diamond$ needs to be chosen larger than $\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}$ to decrease the variance. So, consider two cases:

1. If the noise is small, meaning that

$$v_{\boldsymbol{\varepsilon}}^2 \frac{n\sum_{i>k}\lambda_i^2}{n\lambda_{k+1}\sum_{i>k}\lambda_i} \leq \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}} \frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n},$$

then set

$$\Diamond = a\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}$$

for a constant a that only depends on σ_x that we will choose next. This a must be such that Equation (A.20) is satisfied, which means

$$\begin{aligned} a\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} \leq& \sum_{i>k}\lambda_i, \\ a\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} \geq& 2c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right). \end{aligned}$$

Using $\sqrt{n\sum_{i>k}\lambda_i^2} \leq \sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}$ we obtain that it is enough for a to satisfy

$$\begin{aligned} a \leq& \rho_k(0)^{1/2}, \\ a \geq& 2c_1\left(\rho_k(0)^{-1/2} + 1\right). \end{aligned}$$

One can see that $a = 4c_1$ satisfies this condition when $c > \max(1, 16c_1^2)$ since $\rho_k(0) > c$. Taking such an a, plugging $\Diamond$ into Equations (A.21)–(A.23), and choosing c_4 big enough depending on a, c_1, c_2, c_3, we get that with probability at least $1 - c_4e^{-n/c_4}$,

$$\begin{aligned} B + v_{\boldsymbol{\varepsilon}}^2 V \leq& c_4\left(\frac{k}{n}v_{\boldsymbol{\varepsilon}}^2 + v_{\boldsymbol{\varepsilon}}^2\frac{n\sum_{i>k}\lambda_i^2}{n\lambda_{k+1}\sum_{i>k}\lambda_i} + \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n}\right) \\ \leq& 2c_4\left(\frac{k}{n}v_{\boldsymbol{\varepsilon}}^2 + \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n}\right), \end{aligned}$$

which implies the desired bound for any $c > 2c_4$.

2. If the noise is large, meaning that

$$v_{\boldsymbol{\varepsilon}}^2 \frac{n\sum_{i>k}\lambda_i^2}{n\lambda_{k+1}\sum_{i>k}\lambda_i} > \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}} \frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n}, \tag{A.24}$$

then set

$$\diamondsuit = a\sqrt{\frac{v_{\boldsymbol{\varepsilon}}}{\|\bar{\boldsymbol{\theta}}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}} n\sqrt{n\sum_{i>k}\lambda_i^2}}.$$

for a constant a that only depends on σ_x that we choose next. As in the previous case, a must be such that Equation (A.20) is satisfied, which means

$$\begin{aligned} a\sqrt{\frac{v_{\boldsymbol{\varepsilon}}}{\|\bar{\boldsymbol{\theta}}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}} n\sqrt{n\sum_{i>k}\lambda_i^2}} &\leq \sum_{i>k}\lambda_i, \\ a\sqrt{\frac{v_{\boldsymbol{\varepsilon}}}{\|\bar{\boldsymbol{\theta}}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}} n\sqrt{n\sum_{i>k}\lambda_i^2}} &\geq 2c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right). \end{aligned}$$

The first condition is satisfied whenever $a < \sqrt{c}$ due to Equation (2.18). Now consider the second condition. Because of Equation (A.24), we have

$$v_{\boldsymbol{\varepsilon}}\sqrt{n\sum_{i>k}\lambda_i^2} \geq \|\bar{\boldsymbol{\theta}}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}\lambda_{k+1}\sum_{i>k}\lambda_i, \tag{A.25}$$

$$\frac{\diamondsuit}{a} = \sqrt{\frac{v_{\boldsymbol{\varepsilon}}}{\|\bar{\boldsymbol{\theta}}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}} n\sqrt{n\sum_{i>k}\lambda_i^2}} \geq \sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}. \tag{A.26}$$

Thus, it is enough to satisfy

$$a\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} \geq 2c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

This is exactly the same condition as in the previous case, so it can be reduced to

$$a \geq 2c_1(\rho_k(0)^{-1/2} + 1).$$

Thus, just as in the small variance case, we see that since $c > \max(1, 16c_1^2)$ then $a = 4c_1$ satisfies both conditions.

Take such an a. Before plugging $\diamondsuit$ into Equations (A.21)–(A.23), note the following. Because of Equation (A.26), we have

$$\begin{aligned} \frac{n\lambda_{k+1}\sum_{i>k}\lambda_i}{\diamondsuit^2} &\leq \frac{1}{a^2}, \\ \frac{\diamondsuit^2}{n^2} &\geq a^2\frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n}, \end{aligned}$$

which means that if we take c_5 large enough depending on a and c_3, then Equations (A.21)–(A.23) imply

$$\begin{aligned} B/c_5 \leq & \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + \|\bar{\boldsymbol{\theta}}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \frac{\lozenge^2}{n^2}, \\ V/c_5 \leq & \frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{\lozenge^2}. \end{aligned}$$

Now plugging in the expression for $\lozenge$ gives that with probability at least $1 - c_1 e^{-n/c_1}$,

$$B + v_{\boldsymbol{\varepsilon}}^2 V \leq c_5 \left(\frac{k}{n} v_{\boldsymbol{\varepsilon}}^2 + (a^{-2} + a^2) v_{\boldsymbol{\varepsilon}} \|\bar{\boldsymbol{\theta}}_{0:k}\|_{\boldsymbol{\Sigma}^{-1}_{0:k}} \sqrt{\frac{\sum_{i>k} \lambda_i^2}{n}} + \|\bar{\boldsymbol{\theta}}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \right),$$

which implies the result for $c > \max((a^{-2} + a^2)c_5, c_1)$.

□

Appendix B

Proofs for Chapter 3

B.1 Formulas for the solution

In this section we derive the explicit formula for $\boldsymbol{w}_{\text{ridge}}$, which operates with inverse of matrix $\boldsymbol{A}$ instead of $\boldsymbol{X}\boldsymbol{X}^\top$. The version of this formula for the case of MNI solution $\boldsymbol{w}_{\text{MNI}}$ with clean labels $\hat{\boldsymbol{y}} = \boldsymbol{y}$ already appeared in [8], who, in their turn, took it from [58].

Lemma 88 (Explicit formulas for MNI)**.** *Denote $\Delta\boldsymbol{y} := \hat{\boldsymbol{y}} - \boldsymbol{y}$. Denote the projection of $\boldsymbol{\mu}$ on the orthogonal complement to the span of the columns of $\boldsymbol{Q}^\top$ as $\boldsymbol{\mu}_\perp$, and take $\lambda = 0$. Denote*

$$S = (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \|\boldsymbol{\mu}_\perp\|^2 \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}.$$

Then

$$\begin{aligned}
S\boldsymbol{w}_{MNI} &= \left[(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \|\boldsymbol{\mu}_\perp\|^2 \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right] \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \\
&\quad + \left[(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})(1 - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}) - \|\boldsymbol{\mu}_\perp\|^2 \Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right] \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
&\quad + \left[\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\right] \boldsymbol{\mu}_\perp, \\
S\boldsymbol{\mu}^\top \boldsymbol{w}_{MNI} &= \boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\|\boldsymbol{\mu}_\perp\|^2 + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \\
&= \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
&\quad + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}) + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot \|\boldsymbol{\mu}_\perp\|^2.
\end{aligned}$$

In particular, when $\hat{\boldsymbol{y}} = \boldsymbol{y}$

$$\begin{aligned}
S\boldsymbol{w}_{MNI} &= (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_\perp, \\
S\boldsymbol{\mu}^\top \boldsymbol{w}_{MNI} &= \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}.
\end{aligned}$$

Proof. We defined $\boldsymbol{w}_{\text{MNI}}$ as $\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\hat{\boldsymbol{y}}$. Our goal is to derive a different formula, which would have inverse of $\boldsymbol{Q}\boldsymbol{Q}^\top$ instead of $\boldsymbol{X}\boldsymbol{X}^\top$. This derivation could be made algebraically using the fact that $\boldsymbol{X}\boldsymbol{X}^\top$ is a low-rank correction to $\boldsymbol{Q}\boldsymbol{Q}^\top$ and applying Sherman-Morrison-Woodbury identity. Such derivation would be very bulky, so we take another path here and derive the required formula from scratch using geometric considerations. We are going to use the fact that $\boldsymbol{w}_{\mathrm{MNI}}$ can be equivalently defined as the unique vector $\hat{\boldsymbol{w}}$ that lies in the span of columns of $\boldsymbol{X}^\top$ such that $\boldsymbol{X}\hat{\boldsymbol{w}} = \hat{\boldsymbol{y}}$.

Denote the span of the columns of $\boldsymbol{Q}^\top$ as $\mathcal{Q}$, and the projector onto $\mathcal{Q}$ as $\boldsymbol{P_Q} := \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}$. For any $\boldsymbol{v} \in \mathbb{R}^p$ denote the projection of $\boldsymbol{v}$ on $\mathcal{Q}$ as $\boldsymbol{v}_\|$ and the projection of $\boldsymbol{v}$ on $\mathcal{Q}^\perp$ as $\boldsymbol{v}_\perp$:

$$\boldsymbol{v}_\perp := (\boldsymbol{I}_p - \boldsymbol{P_Q})\boldsymbol{v}, \quad \boldsymbol{v}_\| := \boldsymbol{P_Q}\boldsymbol{v} = \boldsymbol{v} - \boldsymbol{v}_\perp.$$

Consider any vector $\boldsymbol{w}$ in the span of the columns of $\boldsymbol{X}^\top$. The projection of this vector on $\mathcal{Q}^\perp$ must be a scalar multiple of $\boldsymbol{\mu}_\perp$ because the projection of the i^{th} column of $\boldsymbol{X}^\top$ is $y_i\boldsymbol{\mu}_\perp$. That is, $\boldsymbol{w}_\perp = \alpha\boldsymbol{\mu}_\perp$ for some scalar α. Now let's answer the following question: which labels does $\boldsymbol{w}$ give to data points? The part in $\mathcal{Q}$ doesn't interact with $\boldsymbol{\mu}_\perp$ (they are orthogonal) and vice versa, so

$$\begin{aligned}\boldsymbol{X}\boldsymbol{w} =&(\boldsymbol{Q} + \boldsymbol{y}\boldsymbol{\mu}_\|^\top)\boldsymbol{w}_\| + \boldsymbol{y}\boldsymbol{\mu}_\perp^\top\boldsymbol{w}_\perp \\ =&\boldsymbol{Q}\boldsymbol{w}_\| + (\boldsymbol{\mu}_\|^\top\boldsymbol{w}_\| + \alpha\|\boldsymbol{\mu}_\perp\|^2)\boldsymbol{y}\end{aligned}$$

Recall that we want to find the minimum norm interpolator for labels $\hat{\boldsymbol{y}}$, that is such $\hat{\boldsymbol{w}}$ that

$$\boldsymbol{Q}\hat{\boldsymbol{w}}_\| + (\boldsymbol{\mu}_\|^\top\hat{\boldsymbol{w}}_\| + \alpha\|\boldsymbol{\mu}_\perp\|^2)\boldsymbol{y} = \hat{\boldsymbol{y}}.$$

Denote $\beta := \boldsymbol{\mu}_\|^\top\hat{\boldsymbol{w}}_\| + \alpha\|\boldsymbol{\mu}_\perp\|^2$. We see that $\hat{\boldsymbol{w}}_\|$ is such vector in $\mathcal{Q}$ that $\boldsymbol{Q}\hat{\boldsymbol{w}}_\| = \hat{\boldsymbol{y}} - \beta\boldsymbol{y}$. Therefore, it is the minimum norm interpolator of labels $\hat{\boldsymbol{y}} - \beta\boldsymbol{y}$ with the data matrix $\boldsymbol{Q}$ and we can use the formula for MNI to obtain

$$\hat{\boldsymbol{w}}_\| = \boldsymbol{Q}^\top(\underbrace{\boldsymbol{Q}\boldsymbol{Q}^\top}_{\boldsymbol{A}})^{-1}(\hat{\boldsymbol{y}} - \beta\boldsymbol{y}).$$

Thus far, we learned that for some scalars α, β it holds that

$$\begin{aligned}\hat{\boldsymbol{w}}_\perp =&\alpha\boldsymbol{\mu}_\perp, \\ \hat{\boldsymbol{w}}_\| =&\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} - \beta\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}, \\ \beta =&\boldsymbol{\mu}_\|^\top\hat{\boldsymbol{w}}_\| + \alpha\|\boldsymbol{\mu}_\perp\|^2 \\ =&\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} - \beta\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y} + \alpha\|\boldsymbol{\mu}_\perp\|^2.\end{aligned}$$

There is, however, one more condition that we are missing: there is only one pair α, β that satisfies the relation above for which the vector $\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} - \beta\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y} + \alpha\boldsymbol{\mu}_\perp$ lies in the span of the columns of $\boldsymbol{X}^\top$ — the one with the minimal norm. Thus, we arrive to the following optimization problem in α, β:

$$\begin{aligned}\alpha^2\|\boldsymbol{\mu}_\perp\|^2 + \beta^2\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y} - 2\beta\hat{\boldsymbol{y}}^\top\boldsymbol{A}^{-1}\boldsymbol{y} \to \min_{\alpha,\beta}&, \\ \text{s.t. } \beta(1 + \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y}) - \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} = \alpha\|\boldsymbol{\mu}_\perp\|^2&,\end{aligned}$$

where in the first line we wrote the expression for $\|\hat{\boldsymbol{w}}\|^2 = \|\hat{\boldsymbol{w}}_\|\|^2 + \|\hat{\boldsymbol{w}}_\perp\|^2$ and dropped the term $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\|^2$ which doesn't depend on α, β.

To solve this problem we parameterize

$$\beta = t\|\boldsymbol{\mu}_\perp\|^2, \quad \alpha = t(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}) - \|\boldsymbol{\mu}_\perp\|^{-2}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}.$$

The optimization problem becomes to minimize the following quantity in t

$$t^2(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{\mu}_\perp\|^2 - 2t(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} + t^2\|\boldsymbol{\mu}_\perp\|^4\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} - 2t\|\boldsymbol{\mu}_\perp\|^2\hat{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\boldsymbol{y},$$

which is a simple minimization of a quadratic function in one variable. We get

$$\begin{aligned}
t =& \frac{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} + \|\boldsymbol{\mu}_\perp\|^2\hat{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{\mu}_\perp\|^2 + \|\boldsymbol{\mu}_\perp\|^4\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}, \\
\beta =& \frac{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} + \|\boldsymbol{\mu}_\perp\|^2\hat{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \|\boldsymbol{\mu}_\perp\|^2\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}, \\
\alpha =& \frac{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} + \|\boldsymbol{\mu}_\perp\|^2\hat{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{\mu}_\perp\|^2 + \|\boldsymbol{\mu}_\perp\|^4\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}) - \|\boldsymbol{\mu}_\perp\|^{-2}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \\
=& \frac{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\|\boldsymbol{\mu}_\perp\|^2\hat{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{\mu}_\perp\|^2 + \|\boldsymbol{\mu}_\perp\|^4\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}} \\
&- \frac{\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\big((1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \|\boldsymbol{\mu}_\perp\|^2\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\big)}{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{\mu}_\perp\|^2 + \|\boldsymbol{\mu}_\perp\|^4\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}} \\
=& \frac{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\hat{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}}{(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \|\boldsymbol{\mu}_\perp\|^2\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}
\end{aligned}$$

Recall that $\Delta\boldsymbol{y} := \hat{\boldsymbol{y}} - \boldsymbol{y}$. Using this notation

$$\hat{\boldsymbol{w}}_\| = \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} + (1 - \beta)\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}, \quad \hat{\boldsymbol{w}}_\perp = \alpha\boldsymbol{\mu}_\perp.$$

Denote

$$S := (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \|\boldsymbol{\mu}_\perp\|^2\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}.$$

We have

$$\begin{aligned}
S(1 - \beta) =& (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})(1 - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}) - \|\boldsymbol{\mu}_\perp\|^2\boldsymbol{y}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}, \\
S\alpha =& (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\hat{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}.
\end{aligned}$$

which gives the desired formula for $\hat{\boldsymbol{w}}$. When it comes to $\boldsymbol{\mu}^\top \hat{\boldsymbol{w}}$, we directly compute the scalar product using the formula for $\hat{\boldsymbol{w}}$:

$$\begin{aligned}
S\boldsymbol{\mu}^\top \hat{\boldsymbol{w}} = & \left[(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \|\boldsymbol{\mu}_\perp\|^2 \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right] \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \\
& + \left[(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})(1-\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}) - \|\boldsymbol{\mu}_\perp\|^2 \Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right] \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
& + \left[\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\right] \|\boldsymbol{\mu}_\perp\|^2 \\
= & \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot \Bigg[(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \|\boldsymbol{\mu}_\perp\|^2 \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
& \quad - (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} - \|\boldsymbol{\mu}_\perp\|^2 \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\Bigg] \\
& + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot \left[(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\|\boldsymbol{\mu}_\perp\|^2 - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2\right] \\
& + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
= & \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
& + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}) + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot \|\boldsymbol{\mu}_\perp\|^2.
\end{aligned}$$

□

Lemma 89 (Explicit formulas for the ridge solution)**.** *Denote*

$$\begin{aligned}
\Delta\boldsymbol{y} := & \hat{\boldsymbol{y}} - \boldsymbol{y}, \\
\boldsymbol{\mu}_{\underset{\sim}{\perp}} := & (\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu}, \\
S := & (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}} \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}.
\end{aligned}$$

Then for any λ such that the matrix $\boldsymbol{A}$ is PD the following holds:

$$\begin{aligned}
S\boldsymbol{w}_{ridge} = & \left[(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}}\right] \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \\
& + \left[(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})(1-\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}) - \Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}}\right] \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
& + \left[\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\right] \boldsymbol{\mu}_{\underset{\sim}{\perp}}, \\
S\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} = & \boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}} + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \\
= & \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}} + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
& + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}) + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot \boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}}.
\end{aligned}$$

In particular, when $\hat{\boldsymbol{y}} = \boldsymbol{y}$

$$S\boldsymbol{w}_{ridge} = (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_{\underset{\sim}{\perp}}, \tag{B.1}$$

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} = (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}}. \tag{B.2}$$

Proof. First of all, we obtain formulas for the particular case when $\lambda = 0$ in Lemma 88 below. These formulas can be extended to the case of positive λ by a standard trick. Recall the definitions:

$$\begin{aligned}\boldsymbol{w}_{\text{ridge}} &= \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_n)^{-1}\hat{\boldsymbol{y}},\\ \boldsymbol{w}_{\text{MNI}} &= \boldsymbol{X}^\dagger \hat{\boldsymbol{y}} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\hat{\boldsymbol{y}}.\end{aligned}$$

Ridge solution can be obtained from the MNI solution with augmented data, namely denote

$$\check{\boldsymbol{Q}} := [\boldsymbol{Q}, \sqrt{\lambda}\boldsymbol{I}_n], \quad \check{\boldsymbol{\mu}} := \begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{0}_n\end{pmatrix}.$$

and

$$\check{\boldsymbol{X}} := \check{\boldsymbol{Q}} + \boldsymbol{y}\check{\boldsymbol{\mu}}^\top = [\boldsymbol{X}, \sqrt{\lambda}\boldsymbol{I}_n].$$

Now MNI solution for the augmented data becomes

$$\check{\boldsymbol{w}}_{\text{MNI}} = \check{\boldsymbol{X}}^\dagger \hat{\boldsymbol{y}} = \begin{pmatrix}\boldsymbol{X}^\top\\ \sqrt{\lambda}\boldsymbol{I}_n\end{pmatrix}(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I}_n)^{-1}\hat{\boldsymbol{y}},$$

that is, $\boldsymbol{w}_{\text{ridge}}$ is equal to the first p coordinates of $\check{\boldsymbol{w}}_{\text{MNI}}$. Moreover, note that $\boldsymbol{\mu}^\top\boldsymbol{w}_{\text{ridge}} = \check{\boldsymbol{\mu}}^\top\check{\boldsymbol{w}}_{\text{MNI}}$. To apply Lemma 88 and obtain the formula for $\check{\boldsymbol{w}}_{\text{MNI}}$ and $\check{\boldsymbol{\mu}}^\top\check{\boldsymbol{w}}_{\text{MNI}}$ we need to plug in the following objects instead of $\boldsymbol{Q}, \boldsymbol{A}, \boldsymbol{\nu}$, and $\boldsymbol{\mu}_\perp$ correspondingly:

$$\begin{aligned}\check{\boldsymbol{Q}} :=& [\boldsymbol{Q}, \sqrt{\lambda}\boldsymbol{I}_n],\\ \check{\boldsymbol{A}} :=& \check{\boldsymbol{Q}}\check{\boldsymbol{Q}}^\top = \boldsymbol{Q}\boldsymbol{Q}^\top + \lambda\boldsymbol{I}_n = \boldsymbol{A},\\ \check{\boldsymbol{\nu}} :=& \check{\boldsymbol{Q}}\check{\boldsymbol{\mu}} = \boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{\nu},\\ \check{\boldsymbol{\mu}}_\perp :=& (\boldsymbol{I}_{p+n} - \check{\boldsymbol{Q}}^\top\check{\boldsymbol{A}}^{-1}\check{\boldsymbol{Q}})\check{\boldsymbol{\mu}} = \begin{pmatrix}(\boldsymbol{I}_p - \boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu}\\ -\sqrt{\lambda}\boldsymbol{A}^{-1}\boldsymbol{\nu}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}_{\underset{\sim}{\perp}}\\ -\sqrt{\lambda}\boldsymbol{A}^{-1}\boldsymbol{\nu}\end{pmatrix}.\end{aligned}$$

The only thing that is not straightforward to plug in is $\|\check{\boldsymbol{\mu}}_\perp\|^2$, which we derive next:

$$\begin{aligned}\|\check{\boldsymbol{\mu}}_\perp\|^2 =& \|(\boldsymbol{I}_p - \boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu}\|^2 + \lambda\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|^2\\ =& \|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{\nu} + \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{\nu} + \lambda\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}p\boldsymbol{A}^{-1}\boldsymbol{\nu}\\ =& \|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{\nu} + \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\underbrace{(\boldsymbol{Q}\boldsymbol{Q}^\top + \lambda\boldsymbol{I}_n)}_{\boldsymbol{A}}\boldsymbol{A}^{-1}\boldsymbol{\nu}\\ =& \|\boldsymbol{\mu}\|^2 - \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{\nu}\\ =& \boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}}.\end{aligned}$$

Now we can obtain the result for $\lambda \geq 0$: Plugging all those objects in Lemma 88 gives the formulas for $\check{\boldsymbol{w}}_{\text{MNI}}$ and $\check{\boldsymbol{\mu}}^\top\check{\boldsymbol{w}}_{\text{MNI}} = \boldsymbol{\mu}^\top\boldsymbol{w}_{\text{ridge}}$. The formula for $\boldsymbol{w}_{\text{ridge}}$ is then obtained from $\check{\boldsymbol{w}}_{\text{MNI}}$ by trimming the last n coordinates.

Finally, to extend the result to the case of negative λ note that the expressions on the both sides of equations in (B.1) are analytic functions of λ on the domain $\{\lambda \in \mathbb{C} : \Re(\lambda) > -\mu_n(\boldsymbol{Q}\boldsymbol{Q}^\top)\}$. Since those equations hold on $\{\lambda \in \mathbb{R} : \lambda > 0\}$ they coincide on that whole domain, in particular for $\{\lambda \in \mathbb{R} : \lambda > -\mu_n(-\mu_n(\boldsymbol{Q}\boldsymbol{Q}^\top))\}$. □

B.2 General probabilistic results

Lemma 90. *Consider a random variable ξ such that*

$$\eta/2 = \mathbb{P}(\xi = 1) = \mathbb{P}(\xi = -1) = (1 - \mathbb{P}(\xi = 0))/2.$$

Then

$$\begin{aligned} \|\xi\|_{\psi_2} =& 1/\sqrt{\ln(1+1/\eta)} \le 1/\sqrt{\ln\frac{3+\eta^{-1}}{2}}, \\ \|\xi^2 - \eta\|_{\psi_2} \le& 1/\sqrt{\ln\frac{3+\eta^{-1}}{2}}. \end{aligned}$$

Proof. By Definition 1, since ξ is a centered random variable

$$\|\xi\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}\exp(\xi^2/t^2) \le 2\right\}.$$

We write out

$$\mathbb{E}\exp(\xi^2/t^2) = \eta e^{t^{-2}} + (1-\eta) \le 2,$$

and see that it is equivalent to $e^{t^{-2}} \le 1 + 1/\eta$. Thus, $\|\xi\|_{\psi_2} = 1/\sqrt{\ln(1+1/\eta)}$.

Now let's do the same for $\xi^2 - \eta$: we need to find some t such that

$$\mathbb{E}\exp\Big((\xi^2-\eta)^2/t^2\Big) = \eta e^{(1-\eta)^2/t^2} + (1-\eta)e^{\eta^2/t^2} \le 2.$$

Let's find such t that a stronger condition is satisfied, namely

$$\begin{aligned} e^{\eta^2/t^2} \le& \frac{3}{2}, \\ \eta e^{(1-\eta)^2/t^2} \le \eta e^{1/t^2} \le& \frac{1}{2} + \frac{3}{2}\eta. \end{aligned}$$

We take

$$t^{-2} = \min\left(\eta^{-2}\ln\frac{3}{2}, \ln\frac{3+\eta^{-1}}{2}\right).$$

Since $\eta^{-1} \ge 1$, we have

$$\eta^{-2}\ln\frac{3}{2} = \ln(3e^{\eta^{-2}}/2) \ge \ln(3(1+\eta^{-2})/2) \ge \ln(3(1+\eta^{-1})/2),$$

so $\|\xi^2 - \eta\|_{\psi_2} \le t = 1/\sqrt{\ln\frac{3+\eta^{-1}}{2}}$.

Finally, we compare two bounds that we obtained:

$$1 + \eta^{-1} - \frac{3+\eta^{-1}}{2} = \frac{\eta^{-1}-1}{2} \ge 0.$$

We see that $1/\sqrt{\ln(1+1/\eta)} \le 1/\sqrt{\ln\frac{3+\eta^{-1}}{2}}$. □

Lemma 91. *Suppose that $\{\eta_i\}_{i=1}^n$ are i.i.d. centered random variables with sub-Gaussian norms σ. Then for some absolute constant $c > 0$ and any $t > 0$ with probability at least $1 - 2e^{-t^2/c}$*

$$\sqrt{\sum_i \eta_i^2} \le \sigma(\sqrt{n} + t)$$

Proof. We basically repeat the proof of Theorem 3.1.1 from [55], but we don't use the assumption that $\{\eta_i\}_{i=1}^n$ have unit variances.

Without loss of generality we can assume that $\sigma = 1$. Indeed, if $\sigma \neq 1$ we can just work with random variables $\{\eta_i/\sigma\}_{i=1}^n$ instead of $\{\eta_i\}_{i=1}^n$.

Denote $v = \sqrt{\mathbb{E}[\eta_1^2]}$ — standard deviation for $\{\eta_i\}_{i=1}^n$. Recall (or note) that $v \le \sigma \le 1$.

As in the proof of Theorem 3.1.1 from [55], we get that random variables $\{\eta_i^2\}_{i=1}^n$ are sub-Exponential, with sub-Exponential norms bounded by an absolute constant. Applying Bernstein's inequality (Corollary 2.8.3) from [55], we get that for some absolute constant $c > 0$ and any $u \ge 0$ with probability at least $1 - 2\exp(-cn(u \wedge u^2))$

$$n^{-1}\sum_i \eta_i^2 \le v^2 + u \le 1 + u \le (1 + (\sqrt{u} \wedge u))^2.$$

Finally, we make a change of variables: $t = \sqrt{n(u \wedge u^2)} = \sqrt{n}(\sqrt{u} \wedge u)$, and get that with probability at least $1 - 2e^{-ct^2}$

$$\sqrt{n^{-1}\sum_i \eta_i^2} \le 1 + t/\sqrt{n}.$$

□

Lemma 92 (Hanson-Wright inequality)**.** *Suppose $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is a (random) matrix and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a centered vector whose components $\{\varepsilon_i\}_{i=1}^n$ are independent, have variances v^2 and sub-Gaussian norms at most σ. If $\boldsymbol{\varepsilon}$ is independent from $\boldsymbol{M}$, then for some absolute constant c and any $s > 0$*

$$\mathbb{P}\left\{|\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} - v^2 tr(\boldsymbol{M})| > \sigma^2 \max(\sqrt{s}\|\boldsymbol{M}\|_F, s\|\boldsymbol{M}\|)\right\} \le 2\exp\{-s/c\}.$$

Proof. This is basically a rewriting of Theorem 6.2.1 (Hanson-Wright inequality) in [55]. According to that theorem, for some absolute constant c for any $t > 0$,

$$\mathbb{P}\left\{|\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} - \mathbb{E}_\varepsilon \boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon}| \ge t\right\} \le 2\exp\left(-c^{-1}\min\left\{\frac{t^2}{\|\boldsymbol{M}\|_F^2\sigma^4}, \frac{t}{\|\boldsymbol{M}\|\sigma^2}\right\}\right),$$

where $\mathbb{E}_\varepsilon$ denotes expectation over $\boldsymbol{\varepsilon}$.

Since for any i, $\mathbb{E}\varepsilon_i = 0$, and $\mathrm{Var}(\varepsilon_i) = v^2$, we have

$$\mathbb{E}\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} = v^2\mathrm{tr}(\boldsymbol{M}).$$

Plug in $t = \sigma^2 \max(\sqrt{s}\|\boldsymbol{M}\|_F, s\|\boldsymbol{M}\|)$, and note that $\frac{t^2}{\|\boldsymbol{M}\|_F^2\sigma^4} \geq s$ and $\frac{t}{\|\boldsymbol{M}\|\sigma^2} \geq s$:

$$\mathbb{P}\left\{|\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} - v^2\mathrm{tr}(\boldsymbol{M})| \geq \sigma^2 \max(\sqrt{s}\|\boldsymbol{M}\|_F, s\|\boldsymbol{M}\|)\right\} \leq 2\exp\left\{-c^{-1}s\right\}.$$

□

Corollary 93 (Weakened Hanson-Wright for PSD matrices)**.** *In the setting of Lemma 92 assume that $\boldsymbol{M}$ is almost surely PSD. Then for some absolute constant $c > 0$ and any $s > 0$*

$$\mathbb{P}\left\{\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} > c\sigma^2(tr(\boldsymbol{M}) + s\|\boldsymbol{M}\|)\right\} \leq 2\exp\left\{-s/c\right\}.$$

Proof. We just need to transform the result of Lemma 92 using the fact that $\boldsymbol{M}$ is PSD. Note that this fact implies that $\|\boldsymbol{M}\|_F^2 \leq \mathrm{tr}(\boldsymbol{M})\|\boldsymbol{M}\|$ so we obtain that with probability at least $1 - 2\exp\left\{-c_1^{-1}s\right\}$

$$|\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} - v^2\mathrm{tr}(\boldsymbol{M})| \leq \sigma^2 \max(\sqrt{s\|\boldsymbol{M}\|\mathrm{tr}(\boldsymbol{M})}, s\|\boldsymbol{M}\|),$$

where c_1 is the constant from Lemma 92. Now on the same even we can write

$$\begin{aligned}\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} \leq & v\mathrm{tr}(\boldsymbol{M}) + \sigma^2\sqrt{s\|\boldsymbol{M}\|\mathrm{tr}(\boldsymbol{M})} + s\|\boldsymbol{M}\| \\ \leq & \sigma^2(\mathrm{tr}(\boldsymbol{M}) + \sqrt{s\|\boldsymbol{M}\|\mathrm{tr}(\boldsymbol{M})} + s\|\boldsymbol{M}\|) \\ \leq & \frac{3}{2}\sigma^2(\mathrm{tr}(\boldsymbol{M}) + s\|\boldsymbol{M}\|),\end{aligned}$$

where we used the fact that $\sigma \geq v$ (sub-Gaussian norm is greater or equal to variance for any centered distribution) in the second line, and AM-GM inequality $2\sqrt{s\|\boldsymbol{M}\|\mathrm{tr}(\boldsymbol{M})} \leq \mathrm{tr}(\boldsymbol{M}) + s\|\boldsymbol{M}\|$ in the last line.

Taking c large enough depending on c_1 yields the result. □

The following lemma is a restatement of Lemma 84 with a change of notation.

Lemma 94. *Suppose that $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{n\times p}$ is a matrix with i.i.d. isotropic σ-sub-Gaussian rows. Suppose that $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ is a symmetric PSD matrix that is independent of $\tilde{\boldsymbol{Z}}$. Then there exists an absolute constant c such that for any $t > 0$ with probability at least $1 - 6e^{-t/c}$*

$$\|\tilde{\boldsymbol{Z}}\boldsymbol{M}\tilde{\boldsymbol{Z}}^\top\| \leq c\sigma^2\big(\|\boldsymbol{M}\|(t+n) + tr(\boldsymbol{M})\big).$$

Corollary 95. *There exists a constant c that only depends on σ_x such that with probability at least $ce^{-n/c}$*

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

Proof. Note that $\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top = \boldsymbol{Z}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}^2\boldsymbol{Z}_{k:\infty}^\top$, apply Lemma 94 for $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}_{k:\infty}$ and $\boldsymbol{M} = \boldsymbol{\Sigma}_{k:\infty}^2$. □

Lemma 96. *Consider $\boldsymbol{y} \in \{-1, 1\}^n$ — random vector with i.i.d. Rademacher coordinates. Suppose that $\boldsymbol{v} \in \mathbb{R}^n$ is independent from $\boldsymbol{y}$. Then for some absolute constant c with probability at least c^{-1}*

$$|\boldsymbol{v}^\top \boldsymbol{y}| \geq c^{-1}\|\boldsymbol{v}\|.$$

Proof. Since $\boldsymbol{y}$ is a vector with centered independent coordinates that have constant sub-Gaussian norms, the random variable $\xi := \boldsymbol{v}^\top \boldsymbol{y}/\|\boldsymbol{v}\|$ has sub-Gaussian norm at most c_1, where c_1 is an absolute constant.

Thus, for some absolute constant c_2 we and any $t > 0$

$$\mathbb{P}(\xi > t) \leq 2e^{-t^2/c_2}.$$

The idea is to consider variance $\mathbb{E}\xi^2 = 1$. Since the tails of the random variable ξ decay very fast, only a small fraction of that variance can come from the tail, which means that most of it must come from a segment of constant length, from which it is easy to deduce the bound by Markov's inequality.

Formally, we can write for any c_3 and $c_4 > c_3$

$$\begin{aligned}
1 =& \mathbb{E}[\xi^2] \\
=& \int_0^\infty \mathbb{P}(|\xi|^2 > t)\, dt \\
\leq& \int_0^{c_3} 1\, dt + \int_{c_3}^{c_4} \mathbb{P}(|\xi|^2 > t)\, dt + 2\int_{c_4}^\infty e^{-t/c_2}\, dt \\
\leq& c_3 + (c_4 - c_3)\mathbb{P}(\xi^2 > c_3) + 2c_2 e^{-c_4/c_2}.
\end{aligned}$$

We see that

$$\mathbb{P}(\xi^2 > c_3) \geq \frac{1 - c_3 - 2c_2 e^{-c_4/c_2}}{c_4}$$

Taking c_4 to be a large enough absolute constant, and c_3 — small enough, yields the result. □

Lemma 97. *Consider $\boldsymbol{y} \in \{-1, 1\}^n$ — random vector with i.i.d. Rademacher coordinates. Suppose that $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is a matrix that is independent from $\boldsymbol{y}$ and almost surely PSD. Then for some absolute constant c with probability at least c^{-1}*

$$|\boldsymbol{y}^\top \boldsymbol{M} \boldsymbol{y}| \geq c^{-1} tr(\boldsymbol{M}).$$

Proof. First of all, since $\boldsymbol{y}$ has i.i.d. centered coordinates with sub-Gaussian norms bounded by an absolute constant, by Corollary 93 for some absolute constant c_1 and any $s > 0$

$$\mathbb{P}(\boldsymbol{y}^\top \boldsymbol{M} \boldsymbol{y} > c_1(\mathrm{tr}(\boldsymbol{M}) + s\|\boldsymbol{M}\|)) \leq 2e^{-s/c_1}.$$

Denote $\xi = \boldsymbol{y}^\top \boldsymbol{M} \boldsymbol{y}/\mathrm{tr}(\boldsymbol{M})$. Recall that our goal is to to show that $\mathbb{P}(\xi > c^{-1}) > c^{-1}$.

Note that $\|\boldsymbol{M}\| \le \operatorname{tr}(\boldsymbol{M})$ since $\boldsymbol{M}$ is PSD. Thus, it follows from the above that for any $s > 0$

$$\mathbb{P}(\xi > c_1(1+s)) \le 2e^{-s/c_1}.$$

For further convenience we rewrite that as follows: for any $t > c_1$

$$\mathbb{P}(\xi > t) \le 2e^{-(t/c_1-1)/c_1}.$$

Now we follow the same strategy as in the proof of Lemma 96. We write for some small c_2, and large $c_3 > c_1$

$$\begin{aligned}1 =&\mathbb{E}[\xi]\\ =&\int_0^\infty \mathbb{P}(|\xi| > t)\,dt\\ \le&\int_0^{c_2} 1\,dt + \int_{c_2}^{c_3} \mathbb{P}(|\xi| > t)\,dt + 2\int_{c_3}^\infty e^{-(t/c_1-1)/c_1}\,dt\\ \le& c_2 + (c_3 - c_2)\mathbb{P}(\xi > c_2) + 2c_1e^{-(c_3/c_1-1)/c_1}.\end{aligned}$$

$$\mathbb{P}(\xi > c_2) \ge \frac{1 - c_2 - 2c_1e^{-(c_3/c_1-1)/c_1}}{c_3 - c_2}.$$

Taking c_2 to be a small enough constant, and c_3 — large enough, yields the result. □

B.3 Some important relations

Lemma 41 (Relations between the main quantities)**.** *Suppose that*

$$k \le n \quad \textit{and} \quad \Lambda > n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}. \tag{3.12}$$

Then

$$n\Diamond^2 \le N, \quad n\Diamond^2 \le N\sqrt{n\Delta V}, \quad V \le 2, \quad \Delta V \le \frac{3}{n}, \quad \Delta V \le 4V.$$

Proof. For the first inequality, we write

$$\begin{aligned}\Diamond^2 =& n\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} + n^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ \le& n\Lambda^{-2}\lambda_{k+1}\|\boldsymbol{\mu}_{k:\infty}\|^2 + n^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ \le& \Lambda^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2 + n^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ =& \Lambda^{-1}M,\end{aligned}$$

where we used that $\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\right\| \le 1$ in the second line, and $\Lambda > n\lambda_{k+1}$ in the third line. Alternatively, we could use $\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\right\| \le n\Lambda^{-1}\lambda_1$ in the second line to obtain

$$\begin{aligned}\Diamond^2 \le& n\Lambda^{-2}\lambda_{k+1}\|\boldsymbol{\mu}_{k:\infty}\|^2 + n\Lambda^{-1}\lambda_1\cdot n^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ \le& n\Lambda^{-2}\lambda_1\left(\|\boldsymbol{\mu}_{k:\infty}\|^2 + n^{-1}\Lambda\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\right)\\ =& n\Lambda^{-2}\lambda_1 M,\end{aligned}$$

which means that

$$\Diamond^2 \le \Lambda^{-1}M \wedge n\Lambda^{-2}\lambda_1 M = \left(\frac{1}{\sqrt{n}} \wedge \frac{\sqrt{n}\lambda_1}{\Lambda}\right)\sqrt{n}\Lambda^{-1}M \le \Lambda^{-1}M\sqrt{n\Delta V}.$$

Now let's upper bound V:

$$\begin{aligned}V =& n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2}n\sum_{i>k}\lambda_i^2\\ \le& k/n + \Lambda^{-2}n\sum_{i>k}\lambda_i^2\\ \le& 2,\end{aligned}$$

where we used Equation (3.12) to make the second transition, and we used the fact that $\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}$ is a $k\times k$ symmetric matrix, all eigenvalues of which are in $(0,1)$.

When it comes to ΔV, we write

$$n\Delta V \le 1 + \frac{n^2\lambda_{k+1}^2 + n\sum_{i>k}\lambda_i^2}{\Lambda^2} \le 3,$$

where the last transition follows directly from Equation (3.12).

Finally, let's compare V and ΔV. In case $k=0$ we get

$$4V = 4\Lambda^{-2}n\sum_i \lambda_i^2 \ge \Lambda^{-2}\left(2n\lambda_1^2 + \sum_i \lambda_i^2\right) = \Delta V.$$

If $k>0$, we have

$$\begin{aligned}4V =& 4n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right) + 4\Lambda^{-2}n\sum_{i>k}\lambda_i^2\\ \ge& 4n^{-1}\frac{1}{(1+\Lambda n^{-1}\lambda_1^{-1})^2} + \Lambda^{-2}n\lambda_{k+1}^2 + \Lambda^{-2}\sum_{i>k}\lambda_i^2\\ \ge& n^{-1}\frac{1}{(1\vee\Lambda n^{-1}\lambda_1^{-1})^2} + \Lambda^{-2}n\lambda_{k+1}^2 + \Lambda^{-2}\sum_{i>k}\lambda_i^2\\ =& \Delta V.\end{aligned}$$

□

Lemma 47 (Bounds via k^*)**.** *Suppose that*

$$k \le n/2 \quad \textit{and} \quad \Lambda > n\lambda_{k+1}.$$

Define

$$\begin{aligned}
k^* :=& \min\left\{\kappa \in \{0, 1, \dots, k\} : \lambda + \sum_{i>\kappa} \lambda_i \ge n\lambda_{\kappa+1}\right\},\\
\Lambda_* :=& \lambda + \sum_{i>k^*} \lambda_i,\\
V_* :=& \frac{k^*}{n} + \Lambda_*^{-2} n \sum_{i>k^*} \lambda_i^2,\\
\Diamond_*^2 :=& n^{-1} \|\boldsymbol{\mu}_{0:k^*}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k^*}} + n\Lambda_*^{-2} \|\boldsymbol{\mu}_{k^*:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}},\\
N_* :=& \|\boldsymbol{\mu}_{0:k^*}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k^*}} + n\Lambda_*^{-1} \|\boldsymbol{\mu}_{k^*:\infty}\|^2.
\end{aligned}$$

Then

$$2N_* \ge N \ge N_*/2, \quad 2\Diamond_* \ge \Diamond \ge \Diamond_*/2, \quad 4V_* \ge V \ge V_*/4, \quad \Lambda_* \ge \Lambda \ge \Lambda_*/2.$$

Proof. First of all, let's compare Λ and Λ_*. Since $k^* \le k$, we obviously have $\Lambda_* \ge \Lambda$. On the other hand,

$$\begin{aligned}
\Lambda_* =& \lambda + \sum_{i=k^*+1}^{k} \lambda_i + \sum_{i>k} \lambda_i\\
\le& \lambda + (k - k^*)\lambda_{k^*+1} + \sum_{i>k} \lambda_i\\
\le& \frac{k - k^*}{n}\Lambda_* + \Lambda\\
\le& \frac{1}{2}\Lambda_* + \Lambda.
\end{aligned}$$

Therefore, $\Lambda_* \le 2\Lambda$.

Suppose that $k^* \ne 0$ (we will deal with the case $k^* = 0$ separately in the end. Let's show that k^* is the "the place where the transition happens", more precisely $\lambda_i \le n^{-1}\Lambda_*$ for $i > k^*$ and $\lambda_i \ge n^{-1}\Lambda_*$ for $i \le k^*$. Indeed, the first of those inequalities follows from the definition of k^*, and for the second we can write

$$n\lambda_i \ge n\lambda_{k^*} > \lambda + \sum_{i \ge k^*} \lambda_i \ge \Lambda_*,$$

where the second inequality also follows from the definition of k^*. Combining with the fact that $\Lambda \le \Lambda_*$, we also obtain that $\lambda_i \ge n^{-1}\Lambda$ for $i \le k^*$.

Now, let's prove the remaining relations one-by-one.

1. $n\Lambda^{-1}M$ vs $n\Lambda_*^{-1}M_*$.

$$\begin{aligned}
n\Lambda^{-1}M &= \left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2\\
&= \sum_{i=1}^{k}\frac{\mu_i^2}{\lambda_i(1+\lambda_i^{-1}n^{-1}\Lambda)} + n\Lambda^{-1}\sum_{i>k}\mu_i^2\\
&= \sum_{i=1}^{k^*}\frac{\mu_i^2}{\lambda_i+n^{-1}\Lambda} + \sum_{i=k^*+1}^{k}\frac{\mu_i^2}{\lambda_i+n^{-1}\Lambda} + \sum_{i>k}\frac{\mu_i^2}{n^{-1}\Lambda}\\
&\begin{cases}
\ge \sum_{i=1}^{k^*}\frac{\mu_i^2}{2\lambda_i} + \sum_{i=k^*+1}^{k}\frac{\mu_i^2}{2n^{-1}\Lambda_*} + \sum_{i>k}\frac{\mu_i^2}{n^{-1}\Lambda_*},\\
\le \sum_{i=1}^{k^*}\frac{\mu_i^2}{\lambda_i} + \sum_{i=k^*+1}^{k}\frac{\mu_i^2}{n^{-1}\Lambda_*/2} + \sum_{i>k}\frac{\mu_i^2}{n^{-1}\Lambda_*/2},
\end{cases}
\end{aligned}$$

where we plugged in $n^{-1}\Lambda \le \lambda_i$ for $i \le k^*$, $\lambda_i \le n^{-1}\Lambda_*$ for $i > k^*$, and $\Lambda_* \ge \Lambda \ge \Lambda_*/2$ in the last transition.

Since

$$n\Lambda_*^{-1}M_* = \sum_{i=1}^{k^*}\frac{\mu_i^2}{\lambda_i} + \sum_{i>k^*}\frac{\mu_i^2}{n^{-1}\Lambda_*},$$

the above implies that $2n\Lambda_*^{-1}M_* \ge n\Lambda^{-1}M \ge n\Lambda_*^{-1}M_*/2$.

2. $\Diamond$ vs $\Diamond_*$.

$$\begin{aligned}
n\Diamond^2 &= \left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n^2\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\\
&= \sum_{i=1}^{k}\frac{\mu_i^2}{\lambda_i(1+\lambda_i^{-1}n^{-1}\Lambda)^2} + n^2\Lambda^{-2}\sum_{i>k}\lambda_i\mu_i^2\\
&= \sum_{i=1}^{k^*}\frac{\lambda_i\mu_i^2}{(\lambda_i+n^{-1}\Lambda)^2} + \sum_{i=k^*+1}^{k}\frac{\lambda_i\mu_i^2}{(\lambda_i+n^{-1}\Lambda)^2} + \sum_{i>k}\frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda)^2}\\
&\begin{cases}
\ge \sum_{i=1}^{k^*}\frac{\lambda_i\mu_i^2}{(2\lambda_i)^2} + \sum_{i=k^*+1}^{k}\frac{\lambda_i\mu_i^2}{(2n^{-1}\Lambda_*)^2} + \sum_{i>k}\frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda_*)^2},\\
\le \sum_{i=1}^{k^*}\frac{\lambda_i\mu_i^2}{\lambda_i^2} + \sum_{i=k^*+1}^{k}\frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda_*/2)^2} + \sum_{i>k}\frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda_*/2)^2},
\end{cases}
\end{aligned}$$

where we plugged in $n^{-1}\Lambda \le \lambda_i$ for $i \le k^*$, $\lambda_i \le n^{-1}\Lambda_*$ for $i > k^*$, and $\Lambda_* \ge \Lambda \ge \Lambda_*/2$ in the last transition.

Since

$$n\Diamond_*^2 = \sum_{i=1}^{k^*}\frac{\lambda_i\mu_i^2}{\lambda_i^2} + \sum_{i>k^*}\frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda_*)^2},$$

the above implies that $4n\Diamond_*^2 \ge n\Diamond^2 \ge n\Diamond_*^2/4$.

3. V vs V_*.

$$\begin{aligned} V =& n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2} n\sum_{i>k}\lambda_i^2 \\ =& n^{-1}\sum_{i=1}^{k}\frac{1}{(1+\lambda_i^{-1}n^{-1}\Lambda)^2} + \Lambda^{-2}n\sum_{i>k}\lambda_i^2 \\ =& \sum_{i=1}^{k^*}\frac{\lambda_i^2/n}{(\lambda_i + n^{-1}\Lambda)^2} + \sum_{i=k^*+1}^{k}\frac{\lambda_i^2/n}{(\lambda_i + n^{-1}\Lambda)^2} + \sum_{i>k}\frac{\lambda_i^2/n}{(n^{-1}\Lambda)^2} \\ & \begin{cases} \geq \sum_{i=1}^{k^*}\frac{\lambda_i^2/n}{(2\lambda_i)^2} + \sum_{i=k^*+1}^{k}\frac{\lambda_i^2/n}{(2n^{-1}\Lambda_*)^2} + \sum_{i>k}\frac{\lambda_i^2/n}{(n^{-1}\Lambda_*)^2}, \\ \leq \sum_{i=1}^{k^*}\frac{\lambda_i^2/n}{\lambda_i^2} + \sum_{i=k^*+1}^{k}\frac{\lambda_i^2/n}{(n^{-1}\Lambda_*/2)^2} + \sum_{i>k}\frac{\lambda_i^2/n}{(n^{-1}\Lambda_*/2)^2}, \end{cases} \end{aligned}$$

where we plugged in $n^{-1}\Lambda \leq \lambda_i$ for $i \leq k^*$, $\lambda_i \leq n^{-1}\Lambda_*$ for $i > k^*$, and $\Lambda_* \geq \Lambda \geq \Lambda_*/2$ in the last transition.

Since

$$V^* = \sum_{i=1}^{k^*}\frac{\lambda_i^2/n}{\lambda_i^2} + \sum_{i>k^*}\frac{\lambda_i^2/n}{(n^{-1}\Lambda_*)^2}$$

the above implies that $4V_* \geq V \geq V_*/4$.

□

Lemma 48 (Alternative form of the bounds)**.** *Suppose that $k < n$ and $\Lambda > n\lambda_{k+1}$. Denote*

$$N_a := \sum_i \frac{\mu_i^2}{\lambda_i + \Lambda/n}, \quad V_a := \sum_i \frac{\lambda_i^2/n}{(\lambda_i + \Lambda/n)^2}, \quad \Diamond_a{}^2 := \sum_i \frac{\lambda_i \mu_i^2/n}{(\lambda_i + \Lambda/n)^2}.$$

Then

$$N \geq N_a \geq N/2, \quad V \geq V_a \geq V/4, \quad \Diamond^2 \geq \Diamond_a{}^2 \geq \Diamond^2/4.$$

Proof. We prove the relations one-by-one. In the last transition in each display below we use the fact that for $i > k$ we have $(\Lambda/n)^{-1} \leq 2(\lambda_i + \Lambda/n)^{-1}$ to obtain the upper bound.

$$\begin{aligned} n\Lambda^{-1}M &= \left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2 \\ &= \sum_{i=1}^{k}\frac{\mu_i^2}{\lambda_i(1+\lambda_i^{-1}n^{-1}\Lambda)} + n\Lambda^{-1}\sum_{i>k}\mu_i^2 \\ &= \sum_{i=1}^{k}\frac{\mu_i^2}{\lambda_i + \Lambda/n} + \sum_{i>k}\frac{\mu_i^2}{\Lambda/n} \\ & \begin{cases} \geq \sum_i \frac{\mu_i^2}{\lambda_i + \Lambda/n} \\ \leq 2\sum_i \frac{\mu_i^2}{\lambda_i+\Lambda/n}. \end{cases} \end{aligned}$$

$$\begin{aligned}
n\Diamond^2 =& \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n^2\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\
=& \sum_{i=1}^{k} \frac{\mu_i^2}{\lambda_i(1+\lambda_i^{-1}n^{-1}\Lambda)^2} + n^2\Lambda^{-2}\sum_{i>k}\lambda_i\mu_i^2\\
=& \sum_{i=1}^{k} \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2} + \sum_{i>k}\frac{\lambda_i\mu_i^2}{(\Lambda/n)^2}\\
& \begin{cases} \geq \sum_i \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2}\\ \leq 4\sum_i \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2}, \end{cases}
\end{aligned}$$

$$\begin{aligned}
V =& n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2}n\sum_{i>k}\lambda_i^2\\
=& n^{-1}\sum_{i=1}^{k}\frac{1}{(1+\lambda_i^{-1}n^{-1}\Lambda)^2} + \Lambda^{-2}n\sum_{i>k}\lambda_i^2\\
=& \sum_{i=1}^{k}\frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2} + \sum_{i>k}\frac{\lambda_i^2/n}{(\Lambda/n)^2}\\
& \begin{cases} \geq \sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}\\ \leq 4\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}, \end{cases}
\end{aligned}$$

□

B.4 Randomness in labels

Lemma 98 (Factoring out randomness in labels)**.** *There exists an absolute constant c s.t. conditionally on the draw of $\boldsymbol{Q}$ for any $t > 0$ with probability at least $1 - ce^{-t^2/c}$ over the draw of $(\boldsymbol{y}, \hat{\boldsymbol{y}})$ all the following hold:*

1.
$$\max(|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|, |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}|) \leq ct\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|.$$

2.
$$|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}| \leq ct\sigma_\eta\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|.$$

3.
$$|\Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| \leq c\|\boldsymbol{A}^{-1}\|\left(n\eta + t\sigma_\eta(\sqrt{n}+t)\right).$$

4.

$$\begin{aligned}\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \geq &(n - n\eta - ct\sigma_\eta\sqrt{n} - k)\mu_1(\boldsymbol{A}_k)^{-1} \\ &-(n\eta + ct\sigma_\eta\sqrt{n} + ct\sqrt{n} + ct^2)\|\boldsymbol{A}^{-1}\|.\end{aligned}$$

5.

$$n\|\boldsymbol{A}^{-1}\| \geq \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \geq (n-k)\mu_1(\boldsymbol{A}_k)^{-1} - c(t\sqrt{n} + t^2)\|\boldsymbol{A}^{-1}\|.$$

6.

$$\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \leq c\sigma_\eta^2\Big(tr(\boldsymbol{A}^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) + t^2\|\boldsymbol{A}^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\|\Big).$$

7.

$$\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \leq c\Big(tr(\boldsymbol{A}^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) + t^2\|\boldsymbol{A}^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\|\Big).$$

Proof. Throughout the whole proof we will use Lemma 90, which states that sub-Gaussian norms of the components of $\Delta\boldsymbol{y}/2$ are at most σ_η. Recall also that sub-Gaussian norms of the components of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are equal to an absolute constant (to be precise, $1/\sqrt{\ln(2)}$). Each time we use c in this proof it denotes a new absolute constant. In the end we take c large enough, so all the statements hold.

1. $|\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y}|, |\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}|$: the bound follows directly from the fact that $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are sub-Gaussian vectors with sub-Gaussian norms bounded by an absolute constant (see Lemma 3.4.2 in [55]), and both $\boldsymbol{y}, \hat{\boldsymbol{y}}$ are independent from $\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}$.

2. $|\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\Delta\boldsymbol{y}|$: the bound follows in the same way as above from the fact that $\boldsymbol{\Delta y}$ is a sub-Gaussian vector with sub-Gaussian norm at most $c\sigma_\eta$, and $\Delta\boldsymbol{y}$ is independent from $\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}$.

3. $|\Delta\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}|$. Denote $\boldsymbol{y}_c = \boldsymbol{y} + \Delta\boldsymbol{y}/2$ — the vector, whose coordinates corresponding to the clean points are equal to their clean labels, and other coordinates zeroed out. Conditionally on $\Delta\boldsymbol{y}$, $\boldsymbol{y}_c$ is a vector with i.i.d. Rademacher coordinates supported on the complement of the support of $\Delta\boldsymbol{y}$. Since Rademacher R.V's. are sub-Gaussian, we have that for some absolute constant c for any $t > 0$ the following holds with probability at least $1 - 2e^{-t^2/c}$:

$$\begin{aligned}|\Delta\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}| =&\Big|-\Delta\boldsymbol{y}^\top\boldsymbol{A}^{-1}\Delta\boldsymbol{y}/2 + \Delta\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}_c\Big| \\ \leq&\Delta\boldsymbol{y}^\top\boldsymbol{A}^{-1}\Delta\boldsymbol{y}/2 + ct\|\boldsymbol{A}^{-1}\Delta\boldsymbol{y}\| \\ \leq&\|\boldsymbol{A}^{-1}\|(\|\Delta\boldsymbol{y}\|^2/2 + ct\|\Delta\boldsymbol{y}\|).\end{aligned}$$

By Lemma 90 squares of coordinates of $\Delta\boldsymbol{y}/2$ are σ_η-sub-Gaussian with mean η, so by General Hoeffding's inequality (Theorem 2.6.2 in [55]) for some absolute constant c and any $t > 0$ with probability at least $1 - 2e^{-t^2/c}$

$$|\|\Delta\boldsymbol{y}\|^2/4 - n\eta| \leq ct\sigma_\eta\sqrt{n}.$$

We could use this result to bound $|\|\Delta\boldsymbol{y}\|$ as well, but then $\sqrt{\sigma_\eta}$ will appear in the bounds. Instead, we use Lemma 91 to give an alternative bound that also holds with probability $1 - 2e^{-t^2/c}$:

$$\|\Delta\boldsymbol{y}\|/2 \leq \sigma_\eta(\sqrt{n} + t).$$

Combining these bounds yields the result.

4. $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}$: denote $\boldsymbol{S}_N \in \mathbb{R}^{n\times n}$ to be a diagonal matrix, such that $\boldsymbol{S}_N[i,i] = -1$ if the label of the i-th data point is noisy, and $\boldsymbol{S}_N[i,i] = 1$ otherwise.

   The matrix $\boldsymbol{S}_N$ is independent from both $\boldsymbol{y}$ and $\boldsymbol{A}$. Now we can write

   $$\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} = \boldsymbol{y}^\top(\boldsymbol{A}^{-1}\boldsymbol{S}_N)\boldsymbol{y}.$$

   By Lemma 92 (Hanson-Wright inequality), for some absolute constant c for any $t > 0$ with probability at least $1 - 2e^{-t^2/c}$

   $$\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \geq \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{S}_N) - ct\|\boldsymbol{A}^{-1}\boldsymbol{S}_N\|_F - ct^2\|\boldsymbol{A}^{-1}\boldsymbol{S}_N\|$$

   Note that

   $$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{S}_N\| =& \|\boldsymbol{A}^{-1}\|,\\ \|\boldsymbol{A}^{-1}\boldsymbol{S}_N\|_F =& \|\boldsymbol{A}^{-1}\|_F \leq \sqrt{n}\|\boldsymbol{A}^{-1}\|.\end{aligned}$$

   We need to bound the number of noisy data points in order to bound $\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{S}_N)$ from below. The number of noisy data points is equal to

   $$\|\Delta\boldsymbol{y}\|_0 = \|\Delta\boldsymbol{y}\|^2/4 \leq n\eta + ct\sigma_\eta\sqrt{n},$$

   where the last inequality was taken from before, and holds with probability at least $1 - 2e^{-t^2/c}$.

   Recall that the $n-k$ largest eigenvalues of $\boldsymbol{A}^{-1}$ are greater or equal to $\mu_1(\boldsymbol{A}_k)^{-1}$. Thus, with probability at least $1 - 2e^{-t^2/c}$.

   $$\begin{aligned}\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{S}_N) \geq& (n - \|\Delta\boldsymbol{y}\|_0 - k)\mu_1(\boldsymbol{A}_k)^{-1} - \|\Delta\boldsymbol{y}\|_0\|\boldsymbol{A}^{-1}\|,\\ \geq& (n - n\eta - ct\sigma_\eta\sqrt{n} - k)\mu_1(\boldsymbol{A}_k)^{-1} - (n\eta + ct\sigma_\eta\sqrt{n})\|\boldsymbol{A}^{-1}\|\end{aligned}$$

   Combining it with Hanson-Wright, we get that with probability at least $1 - 4e^{-t^2/c}$

   $$\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \geq (n - n\eta - ct\sigma_\eta\sqrt{n} - k)\mu_1(\boldsymbol{A}_k)^{-1} - (n\eta + ct\sigma_\eta\sqrt{n} + ct\sqrt{n} + ct^2)\|\boldsymbol{A}^{-1}\|$$

5. $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$: the inequality $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \le n\|\boldsymbol{A}^{-1}\|$ holds with probability one since $\|\boldsymbol{y}\|^2 = n$ almost surely. When it comes to the lower bound, it is simply a particular case of the result for $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}$ proven above for $\eta = 0$.

6. $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\|^2_{\boldsymbol{\Sigma}}$: the bound is a direct consequence of Corollary 93, applied to $\boldsymbol{M} = \boldsymbol{A}^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}$ and $\boldsymbol{\varepsilon} = \Delta\boldsymbol{y}$.

7. $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|^2_{\boldsymbol{\Sigma}}$: the bound is a direct consequence of Corollary 93, applied to $\boldsymbol{M} = \boldsymbol{A}^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}$ and $\boldsymbol{\varepsilon} = \boldsymbol{y}$.

□

B.5 Algebraic decompositions

The purpose of this section is to provide algebraic decompositions of various terms or bounds on them as given by the following Lemma.

Lemma 99 (Algebraic decompositions)**.** *For any $k < n$ all the following hold almost surely on the event that the matrix $\boldsymbol{A}_k$ is PD:*

1.

$$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\| \le& \frac{\mu_1(\boldsymbol{A}_k)}{\mu_n(\boldsymbol{A}_k)}\frac{\sqrt{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|\\ &+ \mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|.\end{aligned}$$

2.

$$\begin{aligned}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}} \ge& \frac{1}{2}\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\left\|\left(\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ &+ \|\boldsymbol{\mu}_{k:\infty}\|^2 - 9\mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{0:k}^\top\boldsymbol{\mu}_{k:\infty}\|^2.\end{aligned}$$

3.

$$\begin{aligned}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}} \le& 3\mu_1(\boldsymbol{A}_k)\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\left\|\left(\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_1(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ &+ \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2.\end{aligned}$$

4.

$$\begin{aligned}&\|\boldsymbol{\mu}_{\perp}\|_{\boldsymbol{\Sigma}}^2\\
\le&2\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-2}\mu_1(\boldsymbol{A}_k)^2\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\
&+2\frac{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}\boldsymbol{Z}_{0:k}^\top)}{\mu_n(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2\mu_n(\boldsymbol{A}_k)^2}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2\\
&+3\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2+3\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|\mu_n(\boldsymbol{A}_k)^{-2}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2\\
&+3\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|\frac{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2\mu_n(\boldsymbol{A}_k)^2}\\
&\times\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2.\end{aligned}\tag{B.3}$$

5.

$$\begin{aligned}tr(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})\le&\frac{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2\mu_n(\boldsymbol{A}_k)^2}tr\left(\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right)\\
&+\mu_n(\boldsymbol{A}_k)^{-2}tr(\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top).\end{aligned}\tag{B.4}$$

6.

$$\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\|\le\frac{\mu_1(\boldsymbol{A}_k)^2}{\mu_n(\boldsymbol{A}_k)^2}\frac{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2}\wedge\frac{\lambda_1^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_n(\boldsymbol{A}_k)^2}+\frac{\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|}{\mu_n(\boldsymbol{A}_k)^2}.$$

The remainder of Section B.5 gives the proof of Lemma 99.

Techniques and proof strategy

The main tool that we are going to use in this section is the following application of Sherman-Morrison-Woodbury (SMW) identity for the matrix $\boldsymbol{A}^{-1}$:

Lemma 100. *If $\boldsymbol{A}_k$ is invertible, then all the following hold:*

$$\boldsymbol{A}^{-1}=\boldsymbol{A}_k^{-1}-\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1},\tag{B.5}$$

$$\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}=\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1},\tag{B.6}$$

$$\boldsymbol{I}_k-\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}=\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}.\tag{B.7}$$

Proof. Equation (B.5) is a direct application of SMW as $\boldsymbol{A}^{-1}=(\boldsymbol{A}_k+\boldsymbol{Q}_{0:k}\boldsymbol{Q}_{0:k}^\top)^{-1}$. To derive (B.6) we write

$$\begin{aligned}\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}=&\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}-\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\\
=&\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k-\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}-\boldsymbol{I}_k\right)\right)\\
=&\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\end{aligned}$$

Finally, we derive (B.7) from (B.6):

$$\begin{aligned}
&\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{0:k} \\
=&\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k} \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \\
=&\boldsymbol{I}_k + \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} - \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right) \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \\
=& \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1}.
\end{aligned}$$

□

Another algebraic result that we will utilize is as follows:

Lemma 101. *Suppose $\boldsymbol{M} \in \mathbb{R}^{k\times k}$ is a PD matrix such that $\alpha \boldsymbol{I}_k \preceq \boldsymbol{M} \preceq \beta \boldsymbol{I}_k$ for some positive scalars $\alpha < \beta$. Then for any vector $\boldsymbol{u} \in \mathbb{R}^k$*

$$\begin{aligned}
\left\|(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})^{-1}\boldsymbol{u}\right\| \geq& \beta^{-1} \left\|(\alpha^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{u}\right\|, &\text{(B.8)}\\
\left\|(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})^{-1}\boldsymbol{u}\right\| \leq& \alpha^{-1} \left\|(\beta^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{u}\right\|. &\text{(B.9)}
\end{aligned}$$

Moreover,

$$\beta^{-2} tr\left((\alpha^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-2}\right) \leq tr\left((\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})^{-2}\right) \leq \alpha^{-2} tr\left((\beta^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-2}\right). \qquad \text{(B.10)}$$

Proof. Denote $\boldsymbol{v} := (\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})^{-1}\boldsymbol{u}$, $\boldsymbol{v}_\alpha := (\boldsymbol{\Sigma}_{0:k}^{-1} + \alpha\boldsymbol{I}_k)^{-1}\boldsymbol{u}$, and $\boldsymbol{v}_\beta := (\boldsymbol{\Sigma}_{0:k}^{-1} + \beta\boldsymbol{I}_k)^{-1}\boldsymbol{u}$. Then

$$\begin{aligned}
\boldsymbol{v}_\alpha :=&(\boldsymbol{\Sigma}_{0:k}^{-1} + \alpha\boldsymbol{I}_k)^{-1}\boldsymbol{u} \\
=&(\boldsymbol{\Sigma}_{0:k}^{-1} + \alpha\boldsymbol{I}_k)^{-1}(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})\boldsymbol{v} \\
=&\boldsymbol{v} + (\boldsymbol{\Sigma}_{0:k}^{-1} + \alpha\boldsymbol{I}_k)^{-1}(\boldsymbol{M} - \alpha\boldsymbol{I}_k)\boldsymbol{v}
\end{aligned}$$

Thus,

$$\|\boldsymbol{v}_\alpha\| \leq \|\boldsymbol{v}\| \left(1 + \|(\boldsymbol{M} - \alpha\boldsymbol{I}_k)\| \|(\boldsymbol{\Sigma}_{0:k}^{-1} + \alpha\boldsymbol{I}_k)^{-1}\|\right) \leq \|\boldsymbol{v}\| \left(1 + \frac{\beta - \alpha}{\alpha}\right) = \frac{\beta}{\alpha}\|\boldsymbol{v}\|,$$

which yields Equation (B.8). Analogously,

$$\begin{aligned}
\boldsymbol{v}_\beta :=&(\boldsymbol{\Sigma}_{0:k}^{-1} + \beta\boldsymbol{I}_k)^{-1}\boldsymbol{u} \\
=&(\boldsymbol{\Sigma}_{0:k}^{-1} + \beta\boldsymbol{I}_k)^{-1}(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})\boldsymbol{v} \\
=&\boldsymbol{v} + (\boldsymbol{\Sigma}_{0:k}^{-1} + \beta\boldsymbol{I}_k)^{-1}(\boldsymbol{M} - \beta\boldsymbol{I}_k)\boldsymbol{v}
\end{aligned}$$

Thus,

$$\|\boldsymbol{v}_\beta\| \geq \|\boldsymbol{v}\| \left(1 - \|(\boldsymbol{M} - \beta\boldsymbol{I}_k)\| \|(\boldsymbol{\Sigma}_{0:k}^{-1} + \beta\boldsymbol{I}_k)^{-1}\|\right) \geq \|\boldsymbol{v}\| \left(1 - \frac{\beta - \alpha}{\beta}\right) = \frac{\alpha}{\beta}\|\boldsymbol{v}\|,$$

which yields Equation (B.9).

Finally, Equation (B.10) is a direct consequence of Equations (B.8) and (B.9): take $\boldsymbol{g}$ to be an isotropic Gaussian vector in $\mathbb{R}^k$. Equations (B.8) and (B.9) give

$$\beta^{-2}\left\|(\alpha^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k)^{-1}\boldsymbol{g}\right\|^2 \le \left\|(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{M})^{-1}\boldsymbol{g}\right\|^2 \le \alpha^{-2}\left\|(\beta^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k)^{-1}\boldsymbol{g}\right\|^2.$$

Taking expectation over $\boldsymbol{g}$ yields Equation (B.10). □

$\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|$

In this section we derive an upper bound on $\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|$. Recall that

$$\begin{gathered}\boldsymbol{\nu}=\boldsymbol{Q}\boldsymbol{\mu}=\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}+\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty},\\ \|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|\le\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\|+\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|.\end{gathered}$$

We bound those two terms separately. For the first term we use Equation (B.6):

$$\begin{aligned}\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}=&\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k}\\ =&\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\boldsymbol{\Sigma}_{0:k}^{1/2}\left(\boldsymbol{I}_k+\boldsymbol{\Sigma}_{0:k}^{1/2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\boldsymbol{\Sigma}_{0:k}^{1/2}\right)^{-1}\boldsymbol{\mu}_{0:k}\\ =&\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\boldsymbol{\Sigma}_{0:k}^{1/2}\left(\boldsymbol{I}_k+\boldsymbol{\Sigma}_{0:k}^{1/2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\boldsymbol{\Sigma}_{0:k}^{1/2}\right)^{-1}\boldsymbol{\mu}_{0:k}\\ =&\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}.\end{aligned}$$

So,

$$\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\|^2\le\|\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-2}\boldsymbol{Z}_{0:k}\|\left\|\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2.$$

Now we use Lemma 101 together with the observation that

$$\begin{gathered}\|\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-2}\boldsymbol{Z}_{0:k}\|\le\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_n(\boldsymbol{A}_k)^{-2},\\ \mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_1(\boldsymbol{A}_k)^{-1}\boldsymbol{I}_k\preceq\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\preceq\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_n(\boldsymbol{A}_k)^{-1}\boldsymbol{I}_k\end{gathered}$$

to write

$$\begin{aligned}&\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\|^2\\ \le&\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_n(\boldsymbol{A}_k)^{-2}\left(\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_1(\boldsymbol{A}_k)^{-1}\right)^{-2}\\ &\times\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ =&\frac{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2}\frac{\mu_1(\boldsymbol{A}_k)^2}{\mu_n(\boldsymbol{A}_k)^2}\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2.\end{aligned}$$

For the $k:\infty$ part we can just write

$$\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\| \le \|\boldsymbol{A}^{-1}\|\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\| \le \mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|.$$

Overall,

$$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\| \le& \frac{\mu_1(\boldsymbol{A}_k)}{\mu_n(\boldsymbol{A}_k)}\frac{\sqrt{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| \\ &+\mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|.\end{aligned}$$

$\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}}$

Bound from below

In this section we derive a lower bound on $\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}}$. First of all, we write

$$\begin{aligned}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}} =&\boldsymbol{\mu}^\top(\boldsymbol{I}_p-\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu} \\ =&\boldsymbol{\mu}_{0:k}^\top(\boldsymbol{I}_k-\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k} \\ &+\boldsymbol{\mu}_{k:\infty}^\top(\boldsymbol{I}_{p-k}-\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty})\boldsymbol{\mu}_{k:\infty} \\ &-2\boldsymbol{\mu}_{0:k}^\top\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}.\end{aligned}$$

We see that this decomposition has 3 terms: energy in the spiked part, energy in the tail, and the cross term. We expect the positive contribution to come from $\boldsymbol{\mu}_{0:k}^\top(\boldsymbol{I}_k-\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}+\boldsymbol{\mu}_{k:\infty}^\top\boldsymbol{I}_{p-k}\boldsymbol{\mu}_{k:\infty}$, the other terms will be upper bounded in absolute value and subtracted from the lower bound. The last term (the cross term) is a bit tricky, because bounding it separately leads to a potentially vacuous bound. The approach we take here is to bound it in terms of the quantities from the first two terms, and then bounding those quantities.

Due to Equation (B.7) the first term becomes

$$\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k}.$$

Now let's apply a similar transformation to the cross-term: we use Equation (B.6) to write

$$\begin{aligned}&\left|\boldsymbol{\mu}_{0:k}^\top\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right| \\ =&\left|\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right| \\ =&\left|\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1/2}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1/2}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right| \\ \le& w\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k} \\ &+w^{-1}\left\|\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1/2}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|^2 \\ \le& w\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k}+w^{-1}\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2,\end{aligned}$$

where we introduced an arbitrary scalar $w > 0$ when we used AM-GM inequality. In the last line we also used the following fact:

$$\left\|\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1/2}\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1/2}\right\| \le 1.$$

Indeed, $\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}$ is larger than $\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1/2}(\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1/2})^\top$ in the sense of the Loewner order.

We take $w = 0.25$. So far we obtained that

$$\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp} \ge \frac{1}{2}\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k} + \|\boldsymbol{\mu}_{k:\infty}\|^2 - 9\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{0:k}^\top\boldsymbol{\mu}_{k:\infty}\|^2,$$

where we also used

$$\begin{aligned}\boldsymbol{\mu}_{k:\infty}^\top(\boldsymbol{I}_{p-k} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty})\boldsymbol{\mu}_{k:\infty} =& \|\boldsymbol{\mu}_{k:\infty}\|^2 - \boldsymbol{\mu}_{k:\infty}^\top \boldsymbol{Q}_{k:\infty}^\top \underbrace{\boldsymbol{A}^{-1}}_{\preceq \boldsymbol{A}_k^{-1}} \boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\\ \ge& \|\boldsymbol{\mu}_{k:\infty}\|^2 - \|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{0:k}^\top\boldsymbol{\mu}_{k:\infty}\|^2.\end{aligned}$$

Now we just need to bound $\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k}$ from below and $\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{0:k}^\top\boldsymbol{\mu}_{k:\infty}\|^2$ from above. We write

$$\begin{aligned}&\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k}\\ =&\boldsymbol{\mu}_{0:k}^\top\boldsymbol{\Sigma}_{0:k}^{-1/2}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\\ \ge&\boldsymbol{\mu}_{0:k}^\top\boldsymbol{\Sigma}_{0:k}^{-1/2}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \mu_n(\boldsymbol{A}_k)^{-1}\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\\ =&\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\left\|\left(\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2.\end{aligned}$$

For the term $\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{0:k}^\top\boldsymbol{\mu}_{k:\infty}\|^2$ we simply do a norm-times-norm bound:

$$\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{0:k}^\top\boldsymbol{\mu}_{k:\infty}\|^2 \le \mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{0:k}^\top\boldsymbol{\mu}_{k:\infty}\|^2$$

Combining everything together gives the bound.

$$\begin{aligned}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp} \ge& \frac{1}{2}\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\left\|\left(\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ &+ \|\boldsymbol{\mu}_{k:\infty}\|^2 - 9\mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{0:k}^\top\boldsymbol{\mu}_{k:\infty}\|^2\end{aligned}$$

Bound from above

In this section we bound $\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp}$ from above. This is easier than bounding it from below. Indeed, recall the decomposition from the previous section:

$$\begin{aligned}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp} =& \boldsymbol{\mu}_{0:k}^\top(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\\ &+ \boldsymbol{\mu}_{k:\infty}^\top(\boldsymbol{I}_{p-k} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty})\boldsymbol{\mu}_{k:\infty}\\ &- 2\boldsymbol{\mu}_{0:k}^\top\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}.\end{aligned}$$

For the first term we had

$$\boldsymbol{\mu}_{0:k}^\top(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k} = \boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k}.$$

and for the cross-term

$$\begin{aligned}&\left|\boldsymbol{\mu}_{0:k}^\top\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right|\\ \leq& w\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k} + w^{-1}\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2,\end{aligned}$$

for any $w > 0$. Here we will take $w = 1$. For the second term we simply write

$$\boldsymbol{\mu}_{k:\infty}^\top(\boldsymbol{I}_{p-k} - \boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty})\boldsymbol{\mu}_{k:\infty} \leq \|\boldsymbol{\mu}_{k:\infty}\|^2.$$

Combining everything together, we get

$$\begin{aligned}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}} \leq& 3\boldsymbol{\mu}_{0:k}^\top\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k} + \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2\\ =& 3\boldsymbol{\mu}_{0:k}^\top\boldsymbol{\Sigma}_{0:k}^{-1/2}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k} + \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2\\ \leq& 3\mu_1(\boldsymbol{A}_k)\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\left\|\left(\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_1(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ &+ \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2.\end{aligned}$$

$\|\boldsymbol{\mu}_{\underset{\sim}{\perp}}\|_{\boldsymbol{\Sigma}}^2$

The quantity $\|\boldsymbol{\mu}_{\underset{\sim}{\perp}}\|_{\boldsymbol{\Sigma}}^2$ is exactly the bias term from Chapter 2 up to the following change of notation: $\boldsymbol{\mu}$ instead of $\boldsymbol{\theta}^*$, $\boldsymbol{Q}$ instead of $\boldsymbol{X}$. We could, in principle, just borrow an algebraic bound from that chapter (Lemma 87) . However, we would like a bound in a slightly different form, so we do a new derivation here.

As before, we start with the first k components and use Lemma 100:

$$\begin{aligned}&\|[\boldsymbol{\mu}_{\underset{\sim}{\perp}}]_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}^2/2\\ \leq&\left\|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + \left\|\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2\\ =&\left\|\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2\\ &+\left\|\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2\\ =&\left\|\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ &+\left\|\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|^2\\ \leq&\left(\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_1(\boldsymbol{A}_k)^{-1}\right)^{-2}\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ &+\mu_1(\boldsymbol{A}_k)^2\mu_n(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-2}\mu_1(\boldsymbol{Z}_{0:k}\boldsymbol{Z}_{0:k}^\top)\mu_n(\boldsymbol{A}_k)^{-2}\left\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|^2,\end{aligned}$$

where in the last transition we used Lemma 101 and the following observation:

$$\left\| \left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k} \right)^{-1} \right\| \le \mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k})^{-1} \le \mu_1(\boldsymbol{A}_k)\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1}.$$

When it comes to the rest of the components, we write

$$\begin{aligned}
&[\boldsymbol{\mu}_{\underset{\sim}{\perp}}]_{k:\infty} \\
=&\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1} \boldsymbol{Q} \boldsymbol{\mu} \\
=&\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{0:k} \boldsymbol{\mu}_{0:k} \\
=&\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k} \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k} \right)^{-1} \boldsymbol{\mu}_{0:k} \\
=&\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k} \left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k} \right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k},
\end{aligned}$$

which yields

$$\begin{aligned}
&\|[\boldsymbol{\mu}_{\underset{\sim}{\perp}}]_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}} \\
\le&\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|^{1/2} \mu_n(\boldsymbol{A})^{-1} \|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\| \\
&+ \|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|^{1/2} \mu_n(\boldsymbol{A}_k)^{-1} \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{1/2} \left\| \left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k} \right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\| \\
\le&\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|^{1/2} \mu_n(\boldsymbol{A}_k)^{-1} \|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\| \\
&+ \|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|^{1/2} \frac{\mu_n(\boldsymbol{A}_k)^{-1} \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{1/2}}{\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \mu_1(\boldsymbol{A}_k)^{-1}} \\
&\times \left\| \left(\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \mu_n(\boldsymbol{A}_k) \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k \right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|,
\end{aligned}$$

where we used Lemma 101 and the fact that $\mu_n(\boldsymbol{A}) \ge \mu_n(\boldsymbol{A}_k)$ in the last transition. Combining everything together and using the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ yields the final bound.

$\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})$

The quantity $\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})$ is exactly the variance term from Chapter 2: as for the bias term, plug in $\boldsymbol{Q}$ instead of $\boldsymbol{X}$. As before, we could in principle use the algebraic decomposition from Lemma 86, but we make a new derivation because we want to obtain the bound in a different form.

$$
\begin{aligned}
&\operatorname{tr}(\boldsymbol{A}^{-1}[\boldsymbol{Q}_{0:k}, \boldsymbol{Q}_{k:\infty}]\boldsymbol{\Sigma}[\boldsymbol{Q}_{0:k}, \boldsymbol{Q}_{k:\infty}]^\top \boldsymbol{A}^{-1})\\
=&\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\Sigma}_{0:k}\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}) + \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1})\\
=&\operatorname{tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\right)\\
&+ \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1})\\
=&\operatorname{tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\right)\\
&+ \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1})
\end{aligned}
$$

Now let's bound these two terms separately. First, recall that for any PSD matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ the following holds: $\operatorname{tr}(\boldsymbol{M}_1\boldsymbol{M}_2) \le \|\boldsymbol{M}_1\|\operatorname{tr}(\boldsymbol{M}_2)$. We use this to bound the first term as follows:

$$
\begin{aligned}
&\operatorname{tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\right)\\
\le&\mu_n(\boldsymbol{A}_k)^{-2}\operatorname{tr}\left(\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\right)\\
=&\mu_n(\boldsymbol{A}_k)^{-2}\operatorname{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k}\right)\\
\le&\mu_n(\boldsymbol{A}_k)^{-2}\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\operatorname{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\right)\\
\le&\frac{\mu_n(\boldsymbol{A}_k)^{-2}\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2\mu_1(\boldsymbol{A}_k)^{-2}}\operatorname{tr}\left(\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right),
\end{aligned}
$$

where we used Lemma 101 in the last transition.

When it comes to the second term, we simply write

$$
\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}) \le \mu_n(\boldsymbol{A}_k)^{-2}\operatorname{tr}\left(\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\right).
$$

$\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\|$

In Chapter 2 the deviations of the variance term in noise were dealt with in the following way: Hanson-Wright inequality states that a quadratic form $\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a vector with i.i.d. centered sub-Gaussian components concentrates around $\operatorname{tr}(\boldsymbol{M})$, with deviations being composed of a sub-Gaussian tail controlled by $\|\boldsymbol{M}\|_F$ and sub-Exponential tail controlled by $\|\boldsymbol{M}\|$. In Chapter 2 the latter two quantities were bounded as $\|\boldsymbol{M}\|_F^2 \le \operatorname{tr}(\boldsymbol{M})^2$ and $\|\boldsymbol{M}\| \le \operatorname{tr}(\boldsymbol{M})$, so only the bound on the trace of the matrix $\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}$ was required. Instead of making such step, we can bound the spectral norm separately, and then use $\|\boldsymbol{M}\|_F^2 \le \|\boldsymbol{M}\|\operatorname{tr}(\boldsymbol{M})$. This section shows the following:

$$
\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\| \le \frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2} + \frac{\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|}{\mu_n(\boldsymbol{A}_k)^2}.
$$

We bound the operator norm as follows: first, we decompose into two terms

$$\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\Sigma}_{0:k}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}.$$

The second term is straightforward:

$$\begin{aligned}
&\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\|\\
\le&\|\boldsymbol{A}^{-1}\|\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|\|\boldsymbol{A}^{-1}\|\\
=&\frac{\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|}{\mu_n(\boldsymbol{A}_k)^2}.
\end{aligned}$$

For the first term we use Equation (B.6) to write

$$\begin{aligned}
&\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\Sigma}_{0:k}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\|\\
=&\left\|\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\right\|\\
=&\|\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\|\\
\le&\left\|\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\right\|\|\boldsymbol{Z}_{0:k}\boldsymbol{Z}_{0:k}^\top\|\|\boldsymbol{A}_k^{-1}\|^2\\
\le&\left(\lambda_1^2\wedge\left\|\left(\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\right\|\right)\|\boldsymbol{Z}_{0:k}\boldsymbol{Z}_{0:k}^\top\|\|\boldsymbol{A}_k^{-1}\|^2\\
\le&\frac{\mu_1(\boldsymbol{A}_k)^2}{\mu_n(\boldsymbol{A}_k)^2}\frac{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2}\wedge\frac{\lambda_1^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_n(\boldsymbol{A}_k)^2}.
\end{aligned}$$

B.6 Randomness in covariates

Lemma 102 (Randomness in covariates)**.** *Consider some $L > 1$. There exists a constant c that only depends on c_B and L such that the following holds. Denote*

$$\Lambda = \lambda + \sum_{i>k}\lambda_i$$

Assume that $k < n/c$ and

$$\Lambda > cn\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}.$$

Then all the following hold on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$:

1.

$$\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\| \le c\left(n^{-1/2}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| + \Lambda^{-1}\sqrt{n}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}\right);$$

2.

$$c\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}} \ge \frac{\Lambda}{n}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| + \|\boldsymbol{\mu}_{k:\infty}\|^2 \ge \boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}}/c;$$

3.

$$\|\boldsymbol{\mu}_{\underset{\sim}{\perp}}\|_{\boldsymbol{\Sigma}}^2/c \le \Lambda^2 n^{-2}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2+\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2;$$

4.

$$tr(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) \le c\left(n^{-1}tr\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right)+\Lambda^{-2}n\sum_{i>k}\lambda_i^2\right);$$

5.

$$\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\| \le c\left(\frac{1}{n}\wedge\frac{n\lambda_1^2}{\Lambda^2}+\frac{n\lambda_{k+1}^2+\sum_{i>k}\lambda_i^2}{\Lambda^2}\right).$$

Proof. Recall that on $\mathscr{A}_k(L)$ we have $\Lambda/L \le \mu_n(\boldsymbol{A}_k) \le \mu_1(\boldsymbol{A}_k) \le L\Lambda$.

The proof is rather straightforward: we plug the bounds from the definition of the event $\mathscr{B}_k(c_B)$ from Section 3.3, and the bounds on eigenvalues of $\boldsymbol{A}_k$ from the definition of the event $\mathscr{A}_k(L)$ into the result of Lemma 99. Recall that c_B is the constant from the definition of $\mathscr{B}_k(c_B)$.

On $\mathscr{A}_k(L)\cap\mathscr{B}_k(c_B)$ all the following hold:

1.

$$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\| \le& \frac{\mu_1(\boldsymbol{A}_k)}{\mu_n(\boldsymbol{A}_k)}\frac{\sqrt{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|\\&+\mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|\\ \le&\frac{c_B^{3/2}L^2}{\sqrt{n}}\cdot c_BL\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|+c_B^{1/2}L\Lambda^{-1}\sqrt{n}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}.\end{aligned}$$

Note that in the last transition we used that for a positive scalar $a<1$ we can write

$$\left\|\left(a\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|\le a^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|.$$

It is easy to see that this is correct since both matrices $\boldsymbol{I}_k$ and $\boldsymbol{\Sigma}_{0:k}$ are diagonal. We will make such a transition several more times throughout this proof, as well as the following for $b\ge 1$:

$$\left\|\left(b\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|\ge b^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|.$$

2. We start with the lower bound on $\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}}$:

$$\begin{aligned}
\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}} \geq & \frac{1}{2}\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \left\| \left(\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\
& + \|\boldsymbol{\mu}_{k:\infty}\|^2 - 9\mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{0:k}^\top \boldsymbol{\mu}_{k:\infty}\|^2 \\
\geq & \frac{\Lambda}{2Lc_Bn} \cdot \frac{1}{c_BL} \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| \\
& + \|\boldsymbol{\mu}_{k:\infty}\|^2 - 9Lc_B\Lambda^{-1}n\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
\geq & \frac{\Lambda}{2L^2c_B^2n} \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| + \|\boldsymbol{\mu}_{k:\infty}\|^2(1 - 9Lc_Bn\lambda_{k+1}\Lambda^{-1}),
\end{aligned}$$

where in the last line we used that $\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \leq \lambda_{k+1}\|\boldsymbol{\mu}_{k:\infty}\|^2$. Note that if we take $c > 18c_BL$ in the end, then $1-9Lc_Bn\lambda_{k+1}\Lambda^{-1} > 0.5$ since we assumed that $\Lambda > cn\lambda_{k+1}$.

Now, we do the upper bound:

$$\begin{aligned}
\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\underset{\sim}{\perp}} \leq & 3\mu_1(\boldsymbol{A}_k)\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \left\| \left(\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1}\mu_1(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\
& + \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \\
\leq & 3Lc_B\Lambda n^{-1} \cdot c_BL \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\
& + \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2Lc_B\Lambda^{-1}n\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
\leq & 3L^2c_B^2\Lambda n^{-1} \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + \|\boldsymbol{\mu}_{k:\infty}\|^2(1 + 2Lc_B\lambda_{k+1}\Lambda^{-1}n),
\end{aligned}$$

where, as before, we used $\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \leq \lambda_{k+1}\|\boldsymbol{\mu}_{k:\infty}\|^2$ in the last line. Note that here $\Lambda > cn\lambda_{k+1}$ implies that $2Lc_B\lambda_{k+1}\Lambda^{-1}n \leq 2Lc_B/c < 1$ for c large enough.

3. The upper bound on $\|\boldsymbol{\mu}_{\underset{\sim}{\perp}}\|_{\boldsymbol{\Sigma}}^2$, is very similar to the bound on the bias term in Chapter 2, but has a slightly different form. We derive it below.

$$
\begin{aligned}
&\|\boldsymbol{\mu}_{\underset{\sim}{\perp}}\|_{\boldsymbol{\Sigma}}^2 \\
\leq& 2\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-2}\mu_1(\boldsymbol{A}_k)^2 \left\| \left(\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\
&+ 2\frac{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}\boldsymbol{Z}_{0:k}^\top)}{\mu_n(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^2\mu_n(\boldsymbol{A}_k)^2}\left\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|^2 \\
&+ 3\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 + 3\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|\mu_n(\boldsymbol{A}_k)^{-2}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \\
&+ 3\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|\frac{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^2\mu_n(\boldsymbol{A}_k)^2} \\
&\times \left\| \left(\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\
\leq& 2c_B^2L^2\Lambda^2n^{-2}\cdot c_B^2L^2\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\
&+ 2L^4c_B^3n^{-1}\cdot c_Bn\left\|\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
&+ 3\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 + 3c_B\left(n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2\right)\cdot L^2\Lambda^{-2}c_Bn\left\|\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
&+ 3c_B\left(n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2\right)\cdot L^4c_B^3n^{-1}\cdot c_B^2L^2\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2,
\end{aligned}
$$

$$
\begin{aligned}
&\|\boldsymbol{\mu}_{\underset{\sim}{\perp}}\|_{\boldsymbol{\Sigma}}^2 \\
\leq& \left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\cdot\left(2c_B^4L^4\Lambda^2n^{-2} + 3c_B^6L^6n^{-1}\left(n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2\right)\right) \\
&+ \left\|\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\cdot\left(2L^4c_B^4 + 3 + 3c_B^2L^2\Lambda^{-2}n\left(n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2\right)\right).
\end{aligned}
$$

Recall that we imposed the assumption that

$$
\Lambda > n\lambda_{k+1}, \quad \Lambda > \sqrt{n\sum_{i>k}\lambda_i^2}.
$$

Therefore

$$
n\left(n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2\right) \leq 2\Lambda^2.
$$

Plugging this inequality in and taking c large enough depending on c_B and L gives the bound.

4. The upper bound on $\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})$, is also very similar to the one derived in Chapter 2, where it is exactly the the variance term, but has a slightly different form. We derive it below:

$$\begin{aligned}\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) \leq&\frac{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2\mu_n(\boldsymbol{A}_k)^2}\operatorname{tr}\left(\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right)\\ &+\mu_n(\boldsymbol{A}_k)^{-2}\operatorname{tr}(\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top)\\ \leq&L^4c_B^3n^{-1}\cdot c_BL\operatorname{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right)+L^2c_B\Lambda^{-2}n\sum_{i>k}\lambda_i^2.\end{aligned}$$

5.

$$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\| \leq&\frac{\mu_1(\boldsymbol{A}_k)^2}{\mu_n(\boldsymbol{A}_k)^2}\frac{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2}\wedge\frac{\lambda_1^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_n(\boldsymbol{A}_k)^2}+\frac{\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|}{\mu_n(\boldsymbol{A}_k)^2}\\ \leq&\frac{L^4c_B^3}{n}\wedge\frac{\lambda_1^2c_Bn}{L^{-2}\Lambda^2}+c_BL^2\Lambda^{-2}\left(n\lambda_{k+1}^2+\sum_{i>k}\lambda_i^2\right).\end{aligned}$$

In all the cases above we see that taking c large enough depending on c_B and L yields the result. □

B.7 Proof of the main lower bound

First of all, we combine Lemma 98 with Lemma 102 to obtain high probability bounds on all the terms that appear in quantities of interest. The result is given by the following

Lemma 103 (High probability bounds on separate terms)**.** *Consider some $L > 1$. There exists a constant c that only depends on c_B and L and an absolute constant c_y such that the following holds. Assume that $\eta < c^{-1}$. Assume that $k < n/c$*

$$\Lambda > cn\lambda_{k+1}\vee\sqrt{n\sum_{i>k}\lambda_i^2}.$$

For any $t\in(0,\sqrt{n}/c_y)$, conditionally on the event $\mathscr{A}_k(L)\cap\mathscr{B}_k(c_B)$, with probability is at least $1-c_ye^{-t^2/2}$ over the draw of $(\boldsymbol{y},\hat{\boldsymbol{y}})$ all the following hold:

1.

$$|\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y}|\vee|\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}| \leq ct\Diamond,$$

2.

$$|\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\Delta\boldsymbol{y}| \leq ct\sigma_\eta\Diamond,$$

3.

$$|\Delta \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| \leq c\sigma_\eta n\Lambda^{-1},$$

4.

$$\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \geq c^{-1}n\Lambda^{-1},$$

5.

$$cn\Lambda^{-1} \geq \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \geq c^{-1}n\Lambda^{-1},$$

6.

$$\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \leq c\sigma_\eta^2(V + t^2\Delta V),$$

7.

$$\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \leq c(V + t^2\Delta V),$$

8.

$$cM \geq \boldsymbol{\mu}^\top \boldsymbol{\mu}_{\utilde{\perp}} \geq c^{-1}M,$$

9.

$$\|\boldsymbol{\mu}_{\utilde{\perp}}\|_{\boldsymbol{\Sigma}} \leq c\Lambda\Diamond/\sqrt{n},$$

Proof. Parts 1, 2, 6 and 7 can be obtained directly from the corresponding parts of Lemma 98 by plugging in the bounds from Lemma 102. Parts 8 and 9 are exactly parts 2 and 3 of Lemma 102. Thus, only parts 3, 4, and 5 require additional explanation, which we provide below.

First of all, note that $\|\boldsymbol{A}^{-1}\| \leq \|\boldsymbol{A}_k^{-1}\| = \mu_n(\boldsymbol{A}_k)^{-1}$, since $\boldsymbol{A}$ is larger than $\boldsymbol{A}_k$ w.r.t. Loewner order. Thus, on $\mathscr{A}_k(L)$ we have

$$\|\boldsymbol{A}^{-1}\| \leq L\Lambda^{-1}, \quad \mu_1(\boldsymbol{A}_k)^{-1} \geq L^{-1}\Lambda^{-1},$$

Now let's explain parts 3, 4, 5 starting with the corresponding parts of Lemma 98. Denote the (absolute) constant from that lemma as c_1. In all the following we plug in the bounds on eigenvalues of $\boldsymbol{A}_k$, together with $\eta < 1/c$, $k < n/c$ and $t < \sqrt{n}/c$. In the end of each derivation we need to take c large enough depending on L and c_1.

By Lemma 98 for every $t \in (0, \sqrt{n}/c_1)$ on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$ we have with probability at least $1 - c_1e^{-t^2/2}$ over the draw of $(\boldsymbol{y}, \hat{\boldsymbol{y}})$

3.

$$\begin{aligned}|\Delta \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| \leq & c_1\|\boldsymbol{A}^{-1}\|\left(n\eta + t\sigma_\eta(\sqrt{n}+t)\right)\\ \leq & c_1 L\Lambda(n\sigma_\eta + \sqrt{n}\sigma_\eta(\sqrt{n}+\sqrt{n})\\ \leq & cn\Lambda^{-1}\sigma_\eta.\end{aligned}$$

4.

$$\begin{aligned}\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \geq & (n - n\eta - c_1 t\sigma_\eta\sqrt{n} - k)\mu_1(\boldsymbol{A}_k)^{-1}\\ & -(n\eta + c_1 t\sigma_\eta\sqrt{n} + c_1 t\sqrt{n} + c_1 t^2)\|\boldsymbol{A}^{-1}\|\\ \geq & L^{-1}\Lambda^{-1}n(1 - 1/c - c_1\sigma_\eta/c - 1/c - L^2/c - c_1L^2\sigma_\eta/c - L^2c_1/c)\\ \geq & c^{-1}n\Lambda^{-1}.\end{aligned}$$

5.

$$cn\Lambda^{-1} \geq nL\Lambda^{-1} \geq n\|\boldsymbol{A}^{-1}\| \geq \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y},$$

and

$$\begin{aligned}\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \geq & (n-k)\mu_1(\boldsymbol{A}_k)^{-1} - c_1(t\sqrt{n}+t^2)\|\boldsymbol{A}^{-1}\|\\ \geq & (n - n/c)L^{-1}\Lambda^{-1} - c_1(n/c + n/c^2)L\Lambda^{-1}\\ \geq & c^{-1}n\Lambda^{-1}.\end{aligned}$$

□

Theorem 42 (Main lower bound)**.** *For any $c_B > 0, L > 1$ there exists a constant c that only depends on c_B and L, such that the following holds. Assume that $\eta < c^{-1}$, $k < n/c$, and*

$$\Lambda > cn\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}.$$

For any $t \in (0, \sqrt{n}/c)$, conditionally on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$, with probability at least $1 - ce^{-t^2/2}$ over the draw of $(\boldsymbol{y}, \hat{\boldsymbol{y}})$, the following inequalities hold for a certain scalar $S > 0$:

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} \geq c^{-1}N - ct\Diamond, \tag{3.13}$$

$$S\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}} \leq c\left(\left[1 + N\sigma_\eta\right]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n}\right). \tag{3.14}$$

That is, if $N > 2c^2t\Diamond$, then on the same event,

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} \geq \frac{1}{2c^2}\frac{N}{\left[1 + N\sigma_\eta\right]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n}}.$$

Proof. Note that increasing c only makes the statement weaker. From the very beginning let's put c to be large enough, so that $c > 1$ and $\sigma_\eta < 1$.

Recall that

$$\Delta V := \frac{k \wedge 1}{n} + \frac{n\lambda_{k+1}^2 + \sum_{i>k} \lambda_i^2}{\left(\lambda + \sum_{i>k} \lambda_i\right)^2}.$$

The plan is to plug in the bounds from Lemma 103 into quantities of interest, the formulas for which are given by Lemma 89. First of all, however, we need to make sure that $S > 0$. This is required to write $\|S\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}} = S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ and then cancel S in the numerator and denominator. This is indeed the case since $S = (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp}$, and in Lemma 103 we bound $\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp}$ and $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ from below by strictly positive quantities.

Let's plug in the bounds Lemma 103 in the formulas from Lemma 89: denote the constant from Lemma 103 as c_1 and write

1. $S\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$: recall that $\boldsymbol{y}_C$ is the vector of labels of clean points: $\boldsymbol{y}_C = \boldsymbol{y} + \Delta\boldsymbol{y}/2 = \hat{\boldsymbol{y}} - \Delta\boldsymbol{y}/2$. Now we write

$$\begin{aligned}
&\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp} + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\\
=&\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp} + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}(\boldsymbol{y}_C - \Delta\boldsymbol{y}/2)\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}(\boldsymbol{y}_C + \Delta\boldsymbol{y}/2)\\
=&\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp} + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} - (\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y})^2/4 + (\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}_C)^2\\
\geq&\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp} + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} - (\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y})^2/4\\
\geq&\frac{n}{c_1^2\Lambda}M - c_1 t\Diamond - c_1^2 t^2 \sigma_\eta^2 \Diamond^2\\
=&\frac{N}{c_1^2} - c_1 t\Diamond - c_1^2 t^2 \sigma_\eta^2 \Diamond^2.
\end{aligned}$$

Recall that by Lemma 41 we have $N \geq n\Diamond^2$, which yields

$$\frac{N}{c_1^2} - c_1^2 t^2 \sigma_\eta^2 \Diamond^2 \geq \frac{N}{c_1^2} - \frac{c_1^2}{c^2} n\Diamond^2 \geq \frac{N}{2c_1^2},$$

where the last transition is correct if c is taken large enough depending on c_1. Thus, we get

$$\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp} + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \geq \frac{N}{2c_1^2} - c_1 t\Diamond.$$

2. $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, the first term:

$$\begin{aligned}
&\left[(1 + |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|)^2 + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp}\right] \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\|_{\boldsymbol{\Sigma}}\\
\leq&\left[(1 + c_1 t\Diamond)^2 + c_1^2 M n\Lambda^{-1}\right]\sqrt{c_1\sigma_\eta^2(V + t^2\Delta V)}\\
\leq&c_1^{2.5}\left[(1 + t\Diamond)^2 + n\Lambda^{-1}M\right]\sigma_\eta\sqrt{V + t^2\Delta V}.
\end{aligned}$$

3. $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, the second term

$$\begin{aligned}
&\left[(1+|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|)(1+|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}|)+|\Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\underset{\sim}{\perp}}\right]\|\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}\\
\le&\left[(1+c_1t\Diamond)(1+c_1\sigma_\eta t\Diamond)+c_1^2M\sigma_\eta n\Lambda^{-1}\right]\sqrt{c_1(V+t^2\Delta V)}\\
\le& c_1^{2.5}\left[1+(1+\sigma_\eta)t\Diamond+\sigma_\eta t^2\Diamond^2+n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V+t^2\Delta V}\\
\le& 2c_1^3\left[1+t\Diamond+n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V+t^2\Delta V},
\end{aligned}$$

where we used that $\sigma_\eta<1$ and $t^2\Diamond^2<n\Diamond^2<n\Lambda^{-1}M$ in the last transition (by Lemma 41).

4. $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, the third term

$$\begin{aligned}
&\left[\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}+(1+|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|)|\Delta\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}|+\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}|\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\Delta\boldsymbol{y}|\right]\|\boldsymbol{\mu}_{\underset{\sim}{\perp}}\|_{\boldsymbol{\Sigma}}\\
\le&\left[\frac{c_1n}{\Lambda}+(1+c_1t\Diamond)\frac{c_1n\sigma_\eta}{\Lambda}+\frac{c_1^2nt\sigma_\eta\Diamond}{\Lambda}\right]\frac{c_1\Lambda}{\sqrt{n}}\Diamond\\
\le& 2c_1^3\sqrt{n}(1+t\sigma_\eta\Diamond)\Diamond,
\end{aligned}$$

where we used $c_1n\Lambda^{-1}(1+\sigma_\eta)\le 2c_1n\Lambda^{-1}$ in the last line to reduce the number of terms.

Combining all the terms for $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, we get that for some new constant c_2 that only depends on L and c_B under the condition that $t\le\sqrt{n}/c_2$ and $\eta<1/c_2$

$$\begin{aligned}
S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}/c_2\le&\left[(1+t\Diamond)^2+n\Lambda^{-1}M\right]\sigma_\eta\sqrt{V+t^2\Delta V}\\
&+\left[1+t\Diamond+n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V+t^2\Delta V}\\
&+\sqrt{n}(1+t\sigma_\eta\Diamond)\Diamond\\
=&\Diamond^2\cdot t\sigma_\eta(\sqrt{n}+t\sqrt{V+t^2\Delta V})\\
&+\Diamond\cdot\left(\sqrt{n}+t(1+2\sigma_\eta)\sqrt{V+t^2\Delta V}\right)\\
&+\left[1+\sigma_\eta+2n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V+t^2\Delta V}
\end{aligned}$$

By Lemma 41

$$V\le 2,\quad \Delta V\le 3/n,\quad t^2\Delta V\le 3.$$

This allows us to obtain the final bound on $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$: plug in the following inequalities:

$$\begin{aligned}
\sqrt{n}+t\sqrt{V+t^2\Delta V}\le&(1+\sqrt{5})\sqrt{n},\\
\sqrt{n}+t(1+2\sigma_\eta)\sqrt{V+t^2\Delta V}\le&(1+3\sqrt{5})\sqrt{n},\\
1+\sigma_\eta\le&2.
\end{aligned}$$

We get for some c_3 that only depends on L, c_B:

$$S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}} \le c_3 \left(\left[1 + n\Lambda^{-1}M\sigma_\eta\right] \sqrt{V + t^2\Delta V} + \Diamond\sqrt{n} + t\sigma_\eta \Diamond^2 \sqrt{n} \right). \tag{B.11}$$

Finally, note that by Lemma 41 we have $\Diamond^2 \le \Lambda^{-1}M\sqrt{n\Delta V}$, so

$$t\sigma_\eta\Diamond^2\sqrt{n} \le t\sigma_\eta M\Lambda^{-1}n\sqrt{\Delta V} = \sigma_\eta M\Lambda^{-1}n\sqrt{t^2\Delta V} \le n\Lambda^{-1}M\sigma_\eta\sqrt{V + t^2\Delta V}.$$

We see that the term $t\sigma_\eta\Diamond^2\sqrt{n}$ is dominated by another term up to a constant factor, so it can be removed. This gives the final form of the bound.

□

B.8 Proof of tightness

The goal of this section is to prove a constant probability upper bound on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ for the case without label flipping noise (that is, $\boldsymbol{y} = \hat{\boldsymbol{y}}$). We are going to do it by separately bounding $S\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ from above and $\|S\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ from below, where S is the scalar from Lemma 89. With our techniques, the bounds from below are usually more complicated then the bounds from above. This happens because of the cross-terms: one can use Cauchy-Schwarz to bound them from above, but not from below. To overcome this issue, we introduce two additional random signs to the data. More precisely, introduce two independent Rademacher random variables ε_y and ε_q, which are independent from $\boldsymbol{y}$ and $\boldsymbol{Q}$, and denote

$$\bar{\boldsymbol{Q}} := [\boldsymbol{Q}_{0:k}, \varepsilon_q \boldsymbol{Q}_{k:\infty}], \tag{B.12}$$

$$\bar{\boldsymbol{y}} := \varepsilon_y \boldsymbol{y}, \tag{B.13}$$

$$\bar{\boldsymbol{w}}_{\text{ridge}} := (\bar{\boldsymbol{Q}} + \bar{\boldsymbol{y}}\boldsymbol{\mu}^\top)^\top (\underbrace{\bar{\boldsymbol{Q}}\bar{\boldsymbol{Q}}^\top + \lambda \boldsymbol{I}_n}_{=\boldsymbol{A}})^{-1} \bar{\boldsymbol{y}}. \tag{B.14}$$

Note that since the distribution of $\boldsymbol{y}$ is symmetric (i.e. $\boldsymbol{y}$ and $-\boldsymbol{y}$ have the same distribution), the distribution of $\bar{\boldsymbol{y}}$ is the same as the distribution of $\boldsymbol{y}$. We are also going to assume that $\boldsymbol{Q}_{k:\infty}$ is independent from $\boldsymbol{Q}_{0:k}$ and that the distribution of $\boldsymbol{Q}_{k:\infty}$ is symmetric, which implies that the $\boldsymbol{Q}$ is has the same distribution as $\bar{\boldsymbol{Q}}$. Moreover, note that the expressions in the definitions of the events $\mathscr{B}_k(c_B)$ and $\mathscr{A}_k(L)$ don't change if we substitute $\boldsymbol{Q}$ by $\bar{\boldsymbol{Q}}$ in those definitions. Both random signs ε_y and ε_y cancel. For example, this implies Lemma 102 applies if we substitute $\boldsymbol{Q}$ by $\bar{\boldsymbol{Q}}$, and its result holds almost surely over ε_q.

Introduction of those random signs allows us to say that the cross terms are non-negative with probability 0.5 independently of $\boldsymbol{Q}$ and $\boldsymbol{y}$, and thus we don't need to lower bound them to obtain results with constant probability.

By Lemma 89

$$\begin{aligned}\bar{S} :=& (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})^2 + \boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\perp} \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}},\\ \bar{S}\bar{\boldsymbol{w}}_{\text{ridge}} =& (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\bar{\boldsymbol{\mu}}_{\perp},\\ \bar{S}\boldsymbol{\mu}^\top\bar{\boldsymbol{w}}_{\text{ridge}} =& \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\boldsymbol{\mu}^\top\bar{\boldsymbol{\mu}}_{\perp} + (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}},\end{aligned}$$

where we introduced $\bar{\boldsymbol{\nu}} := \bar{\boldsymbol{Q}}\boldsymbol{\mu}$ and $\bar{\boldsymbol{\mu}}_{\perp} = (\boldsymbol{I}_p - \bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{Q}})\boldsymbol{\mu}$.

The remainder of this section is organized as follows: in Section B.8 we bound $\bar{S}\boldsymbol{\mu}^\top\bar{\boldsymbol{w}}_{\text{ridge}}$ from above, in Section B.8 we bound $\|\bar{S}\boldsymbol{\mu}^\top\bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ from below. In Section B.8 we combine those bouds into the upper bound on $\bar{S}\boldsymbol{\mu}^\top\bar{\boldsymbol{w}}_{\text{ridge}}/|\bar{S}\boldsymbol{\mu}^\top\bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, and thus $\boldsymbol{\mu}^\top\boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ too because it has the same distribution.

Numerator

We start with the following auxiliary lemma, which gives separate bounds on two quantities of interest that arise in the proof of the upper bound on $\bar{S}\boldsymbol{\mu}^\top\bar{\boldsymbol{w}}_{\text{ridge}}$.

Lemma 104. *Suppose that the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. For any $L \geq 1$ there exists a constant c that only depends on L, σ_x and c_B such that the following holds. Suppose that $k < n/c$ and $\boldsymbol{Q}_{0:k}$ is independent from $\boldsymbol{Q}_{k:\infty}$. There exists an event $\mathscr{C}$ whose probability is at least $1 - ce^{-n/c}$ such that all the following hold on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B) \cap \mathscr{C}$:*

$$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\|^2 \geq& c^{-1}n^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2,\\ \|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \geq& c^{-1}\Lambda^{-2}n\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}.\end{aligned}$$

Proof. We prove the inequalities separately. Recall that c_B is the constant from the definition of $\mathscr{B}_k(c_B)$ in Section 3.3.

1. Using the expressions that we derived in Section B.5 we have on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$

$$\begin{aligned}&\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\|^2\\ =&\left\|\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ \geq&\mu_1(\boldsymbol{A}_k)^{-2}\mu_n(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\left\|\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ \geq&\frac{\mu_n(\boldsymbol{A}_k)^2\mu_n(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2}\left\|\left(\mu_1(\boldsymbol{A}_k)\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ \geq&L^4c_B^3n^{-1}\cdot Lc_B\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2.\end{aligned}$$

   where we used Lemma 101 in the penultimate line.

2. Use Sherman-Morrison-Woodbury for the matrix $\boldsymbol{A}^{-1}$ (Equation (B.5)):

$$\boldsymbol{A}^{-1} = (\boldsymbol{A}_k + \boldsymbol{Q}_{0:k}\boldsymbol{Q}_{0:k}^\top)^{-1} = \boldsymbol{A}_k^{-1} - \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k})^{-1}\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}.$$

Let's consider the matrix $\boldsymbol{A}_k\boldsymbol{A}^{-1}$ and see what happens when we multiply it by $\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}$. The idea is to say that the column span of $\boldsymbol{Q}_{0:k}$ is independent of $\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}$, and thus the part that lies that span doesn't influence the norm of the vector much. Formally, denote the projector on the orthogonal complement to the span of the columns of $\boldsymbol{Q}_{0:k}$ as $\boldsymbol{P}_{0:k}^\perp \in \mathbb{R}^{n\times n}$, and note (from Equation (B.5)) that $\boldsymbol{P}_{0:k}^\perp \boldsymbol{A}_k\boldsymbol{A}^{-1} = \boldsymbol{P}_{0:k}^\perp$. We write

$$\begin{aligned}
&\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2\\
\geq&\mu_1(\boldsymbol{A}_k)^{-2}\|\boldsymbol{A}_k\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2\\
\geq&\mu_1(\boldsymbol{A}_k)^{-2}\|\boldsymbol{P}_{0:k}^\perp\boldsymbol{A}_k\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2\\
\geq&\mu_1(\boldsymbol{A}_k)^{-2}\|\boldsymbol{P}_{0:k}^\perp\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2.
\end{aligned}$$

Now note that the vector $\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}$ has i.i.d. components, whose variances are equal to $\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}}^2$ and whose sub-Gaussian constants don't exceed $\sigma_x\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}}$. Moreover, $\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}$ is independent from $\boldsymbol{P}_{0:k}^\perp$, and since $\boldsymbol{P}_{0:k}^\perp$ is a projector, we have

$$\|\boldsymbol{P}_{0:k}^\perp\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 = (\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty})^\top\boldsymbol{P}_{0:k}^\perp\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}.$$

Thus, by Hanson-Wright inequality (Lemma 92) for some absolute constant c_1 and any $s > 0$, with probability at least $1 - 2\exp\{-s/c_1\}$,

$$|\|\boldsymbol{P}_{0:k}^\perp\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 - \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}}^2\operatorname{tr}(\boldsymbol{P}_{0:k}^\perp)| \leq \sigma_x^2\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}}^2\max(\sqrt{s}\|\boldsymbol{P}_{0:k}^\perp\|_F, s\|\boldsymbol{P}_{0:k}^\perp\|).$$

Once again, $\boldsymbol{P}_{0:k}^\perp$ is a projector of rank $n-k$, so

$$\operatorname{tr}(\boldsymbol{P}_{0:k}^\perp) = n-k > n/2, \quad \|\boldsymbol{P}_{0:k}^\perp\|_F = \sqrt{n-k} \leq \sqrt{n}, \quad \|\boldsymbol{P}_{0:k}^\perp\|_F = 1.$$

Taking $s = n/c_2$ for a large enough constant c_2 that only depends on σ_x we see that the probability of the following event is at least $1 - 2\exp\{-n/(c_1c_2)\}$:

$$\mathscr{C} := \left\{\|\boldsymbol{P}_{0:k}^\perp\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 > n\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}}^2\left(\frac{1}{2} - \sigma_x^2/\sqrt{c_2}\right)\right\}.$$

For $c_2 > 16\sigma_x^4$ on$\mathscr{C}$ we have $\|\boldsymbol{P}_{0:k}^\perp\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \geq n\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}}^2/4$.

Combining everything together, on $\mathscr{A}_k(L)\cap\mathscr{C}$ we get

$$\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \geq 0.25L^{-2}\Lambda^{-2}n\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}}^2.$$

□

Our upper bound on $\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}$ is given by the following lemma.

Lemma 105. *Suppose that the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. Consider some $L > 1$. There exist large constants a, c that only depend on σ_x, c_B and L and an absolute constant c_y such that the following holds. Assume that $k < n/c$ and*

$$\Lambda > cn\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}.$$

1. *If $n\Lambda^{-1}M \geq a^{-1}\Diamond$, then on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$ for any $t \in (0, \sqrt{n})$ with probability at least $1 - c_y e^{-t^2/c_y}$ over the draw of $\boldsymbol{y}$ almost surely over the draw of $(\varepsilon_q, \varepsilon_y)$*

$$\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{ridge} < c(1+t)n\Lambda^{-1}M.$$

2. *If $n\Lambda^{-1}M < a^{-1}\Diamond$, there exists an event $\mathscr{C}$ that only depends on $\boldsymbol{Q}$, whose probability is at least $1 - ce^{-n/c}$ such that then on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B) \cap \mathscr{C}$ with probability at least c_y^{-1} over the draw of $\boldsymbol{y}$ and $(\varepsilon_q, \varepsilon_y)$*

$$\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{ridge} < 0.$$

Proof. Recall that

$$\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}} = \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}} + (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}.$$

First of all, denote the constant from Lemma 102 as c_1. By that lemma on $\mathscr{A}_k(L)\cap\mathscr{B}_k(c_B)$ we have

$$\begin{aligned}
\bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}} \leq& \mu_n(\boldsymbol{A})^{-1}\|\bar{\boldsymbol{y}}\|^2\boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}}\\
\leq& L\Lambda^{-1}n \cdot c_1 M;\\
\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\| \leq& c_1\Diamond.
\end{aligned}$$

Recall that we indeed can apply Lemma 102 to $\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\|$ instead of $\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|$ because introducing ε_q into the matrix $\boldsymbol{Q}$ does not change the definitions of the events $\mathscr{A}_k(L)$ and $\mathscr{B}_k(c_B)$.

In the same way as in the proof of Lemma 98 since $\boldsymbol{y}$ is a sub-Gaussian vector with sub-Gaussian norm bounded by an absolute constant, we have for some absolute constant $c_{y,1}$ that for any $t > 0$ on $\mathscr{A}_k(L)$ and $\mathscr{B}_k(c_B)$ with probability at least $1 - c_{y,1}e^{-t^2/c_{y,1}}$ over the draw of $\boldsymbol{y}$ almost surely over the draw of ε_q

$$|\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}| = |\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| \leq c_1 t\Diamond.$$

We can make that statement almost surely over ε_q, because it can only take two values, so we can just do multiplicity correction by adjusting the constant $c_{y,1}$.

These upper bounds directly imply the first part of the lemma. Indeed,

$$\begin{aligned}\bar{S}\boldsymbol{\mu}^\top\bar{\boldsymbol{w}}_{\text{ridge}} =&\bar{\boldsymbol{y}}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\boldsymbol{\mu}^\top(\boldsymbol{I}_p-\bar{\boldsymbol{Q}}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{Q}})+(1+\bar{\boldsymbol{\nu}}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{y}})\bar{\boldsymbol{\nu}}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\\ \le& c_1^2n\Lambda^{-1}M+c_1t\Diamond+(c_1t\Diamond)^2\\ \le& c_1^2n\Lambda^{-1}M+c_1t\Diamond+c_1n\Diamond^2\\ \le& n\Lambda^{-1}M(c_1^2+c_1)+c_1tan\Lambda^{-1}M,\end{aligned}$$

where we used that $t<\sqrt{n}$, $\Diamond^2\le\Lambda^{-1}M$ (Lemma 41) and $\Diamond<an\Lambda^{-1}M$ in the last line. In the end, we just need c to be large enough depending on c_1 and a.

When it comes to the second part, we leave the same bounds for the terms $\bar{\boldsymbol{y}}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\boldsymbol{\mu}^\top(\boldsymbol{I}_p-\bar{\boldsymbol{Q}}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{Q}})$ and $(\bar{\boldsymbol{\nu}}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{y}})^2$, but show that the term $\bar{\boldsymbol{\nu}}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{y}}$ can be negative with large enough magnitude to pull the whole bound in the negative direction.

We take the event $\mathscr{C}$ to be the same as in Lemma 104, by which there exists a constant c_2 that only depends on L,σ_x,c_B such that on $\mathscr{A}_k(L)\cap\mathscr{B}_k(c_B)\cap\mathscr{C}$

$$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\|^2\ge& c_2^{-1}n^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2,\\ \|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2\ge& c_2^{-1}\Lambda^{-2}n\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}.\end{aligned}$$

Now we use the same expressions as we derived in Section B.5 to write

$$\begin{aligned}\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\|^2=&\|\boldsymbol{A}^{-1}(\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}+\varepsilon_q\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty})\|^2\\ =&\left\|\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ &+\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2\\ &+2\varepsilon_q(\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty})^\top\boldsymbol{A}^{-2}(\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k})\\ \ge& c_2^{-1}\Diamond^2/2+2\varepsilon_q(\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty})^\top\boldsymbol{A}^{-2}(\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}).\end{aligned}$$

Note that conditionally on $\mathscr{A}_k(L)\cap\mathscr{B}_k(c_B)\cap\mathscr{C}$ with probability 0.5 over the draw of ε_q the term involving ε_q is non-negative, that is $\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\|\ge(2c_2)^{-1/2}\Diamond$. That statement doesn't involve $\boldsymbol{y}$, so $\boldsymbol{y}$ is still independent of this event. Thus, conditionally on it, by Lemma 96 for an absolute constant $c_{y,2}$ with probability at least $c_{y,2}^{-1}$ over the choice of $\boldsymbol{y}$ we have $|\boldsymbol{y}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}|\ge c_{y,2}^{-1}\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\|$. Moreover, we've seen in the first part of the proof that with probability at least $1-c_{y,1}e^{-n/c_{y,1}}$ over the draw of $\boldsymbol{y}$ $|\boldsymbol{y}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}|\le c_1\sqrt{n}\Diamond$. Finally, with probability 0.5 over ε_y we have $\varepsilon_y\boldsymbol{y}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}=-|\boldsymbol{y}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}|$. Combining everything together (recall that $\varepsilon_q,\varepsilon_y,\boldsymbol{Q}$ and $\boldsymbol{y}$ are independent) we get that on $\mathscr{A}_k(L)\cap\mathscr{B}_k(c_B)\cap\mathscr{C}$ with probability at least $0.25\left(c_{y,2}^{-1}-c_{y,1}e^{-n/c_{y,1}}\right)$ over the draw of $\boldsymbol{y}$ and $(\varepsilon_q,\varepsilon_y)$

$$\begin{aligned}\bar{\boldsymbol{y}}^\top\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\le&-c_{y,2}^{-1}(2c_2)^{-1/2}\Diamond,\\ \bar{S}\boldsymbol{\mu}^\top\bar{\boldsymbol{w}}_{\text{ridge}}\le& n\Lambda^{-1}M(c_1^2+2c_1)-c_{y,2}^{-1}(2c_2)^{-1/2}\Diamond\\ <&0,\end{aligned}$$

where the last transition holds for a large enough depending on $c_1,c_2,c_{y,1},c_{y,2}$ since $n\Lambda^{-1}M<a^{-1}\Diamond$. □

Denominator

The next step is to lower-bound the denominator $\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$. Recall that $\varepsilon_y, \varepsilon_q, \boldsymbol{y}, \boldsymbol{Q}$ are all independent from each other. We factor out the randomness in each of those variables one-by-one, starting with the following

Lemma 106. *With probability at least* 0.25 *over the choice of* $(\varepsilon_y, \varepsilon_q)$ *(that is, conditionally on* $\boldsymbol{y}$ *and* $\boldsymbol{Q}$*)*

$$\begin{aligned}
\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{ridge}\|_{\boldsymbol{\Sigma}}^2 \geq & \frac{1}{2}\left\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 \\
& + \frac{1}{2}(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\left(\left\|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + 0.5\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\right) \\
& - \frac{7}{2}(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\left\|\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
& - 7(\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2.
\end{aligned}$$

Proof. Note that for any vectors $\boldsymbol{u}, \boldsymbol{v}$ of the same dimension the following holds:

$$\begin{aligned}
\|\boldsymbol{u}+\boldsymbol{v}\|^2 = & \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 + 2\boldsymbol{u}^\top \boldsymbol{v} \\
\geq & \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 - 2\left(0.25\|\boldsymbol{u}\|^2 + 4\|\boldsymbol{v}\|^2\right) \\
= & 0.5\|\boldsymbol{u}\|^2 - 7\|\boldsymbol{v}\|^2.
\end{aligned}$$

Thus, we write

$$\|\bar{S}\bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}^2 \geq 0.5\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\bar{\boldsymbol{\mu}}_{\utilde{\perp}}\|_{\boldsymbol{\Sigma}}^2 - 7(\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})^2\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}^2.$$

For the last term note that

$$\left\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\right\|_{\boldsymbol{\Sigma}}^2 = \left\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 = \left\|\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + \left\|\varepsilon_y \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 = \left\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2.$$

Next, we decompose the first term as follows:

$$\begin{aligned}
& \left\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\bar{\boldsymbol{\mu}}_{\utilde{\perp}}\right\|_{\boldsymbol{\Sigma}}^2 \\
= & \left\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 + (\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\left\|\bar{\boldsymbol{\mu}}_{\utilde{\perp}}\right\|_{\boldsymbol{\Sigma}}^2 + 2\varepsilon_y f_1(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}, \varepsilon_q \boldsymbol{y}),
\end{aligned}$$

where $f_1(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}, \varepsilon_q \boldsymbol{y})$ is a cross-term, which doesn't involve ε_y. Recall that for the first term we have $\left\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 = \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2$.

For the second term in that decomposition we go a step further and write

$$
\begin{aligned}
&\|\bar{\boldsymbol{\mu}}_{\underset{\sim}{\perp}}\|_{\boldsymbol{\Sigma}}^2 \\
&= \left\|\boldsymbol{\mu}_{0:k} - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{Q}}\boldsymbol{\mu}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + \left\|\boldsymbol{\mu}_{k:\infty} - \varepsilon_q \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{Q}}\boldsymbol{\mu}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
&= \left\|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + \left\|\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
&+ \left\|\varepsilon_q \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + \left\|\varepsilon_q \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
&+ \varepsilon_q f(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}) \\
&\ge \left\|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + \left\|\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 + \varepsilon_q f_2(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}) \\
&\ge \left\|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + 0.5\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
&\quad - 7\left\|\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 + \varepsilon_q f_2(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}),
\end{aligned}
$$

where $f_2(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu})$ is the cross term, which is independent from ε_q, ε_y and $\boldsymbol{y}$.

The statement of the lemma holds on the following event

$$\{\varepsilon_q f_2(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}) \ge 0, \varepsilon_y f_1(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}, \varepsilon_q \boldsymbol{y}) \ge 0,\}$$

whose probability is at least 0.25 conditionally on $\boldsymbol{y}, \boldsymbol{Q}$ since ε_q and ε_y are independent random signs. □

Note that the last term in the lemma above still depends on ε_q through $\bar{\boldsymbol{\nu}}$ and $\bar{\boldsymbol{Q}}$. This is not a problem since we will bound that term almost surely over the draw of ε_q conditionally on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$ and $\boldsymbol{y}$.

The next step is to obtain lower bounds w.r.t. randomness that comes from $\boldsymbol{y}$. Once again, we don't touch the terms that we subtract yet, and only lower-bound the positive terms.

Lemma 107. *There exists an absolute constant c_y such that for any fixed value of $\boldsymbol{Q}$ for any $t \in (0, \sqrt{n}/c_y)$ with probability at least $c_y^{-1} - c_y e^{-t^2/c_y}$ over the draw of $\boldsymbol{y}$ all the following hold almost surely over the draw of ε_q:*

$$
\begin{aligned}
\left\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 &\ge c_y^{-1} tr(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}), \\
\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} &\ge (n-k)\mu_1(\boldsymbol{A}_k)^{-1} - c_y(t\sqrt{n} + t^2)\|\boldsymbol{A}^{-1}\|, \\
\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} &\le n\|\boldsymbol{A}^{-1}\|, \\
\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 &\le c_y(tr(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}) + t^2\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\|), \\
|\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| &\le c_y t\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\|.
\end{aligned}
$$

Proof. The first inequality is a direct application of Lemma 97, and the remaining were shown as a part of Lemma 98. Note that only the last inequality depends on ε_q, and formally Lemma 98 only shows it for a fixed value of ε_q. However, there are only two possible values of ε_q, so the uniform result can be obtained by straightforward multiplicity correction (the constant from Lemma 98 should be doubled). □

At this point we just need the lower bounds on the quantities that only involve $\boldsymbol{Q}$ and $\boldsymbol{\mu}$. These are done by the following

Lemma 108. *Suppose that the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. For any $L \geq 1$ there exists a constant c that only depends on L, c_B and σ_x such that the following holds. Suppose that $k < n/c$ and $\boldsymbol{Q}_{0:k}$ is independent from $\boldsymbol{Q}_{k:\infty}$. There exists an event $\mathscr{C}$ whose probability is at least $1 - ce^{-n/c}$ such that all the following hold on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B) \cap \mathscr{C}$:*

$$\begin{aligned} tr(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) \geq& c^{-1}V, \\ \|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}^2 \geq& c^{-1}\Lambda^2 n^{-2} \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k} \right\|^2. \end{aligned}$$

Proof. We prove the inequalities separately:

1. We start with the first inequality by writing the following.
$$\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) = \mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\Sigma}_{0:k}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}) + \mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}).$$
For the first term we use the same formula as in Section B.5:
$$\begin{aligned} &\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\Sigma}_{0:k}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}) \\ =&\mathrm{tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\right) \\ \geq&\mu_1(\boldsymbol{A}_k)^{-2}\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\right) \\ \geq&\frac{\mu_n(\boldsymbol{A}_k)^2\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2}\mathrm{tr}\left(\left(\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_1(\boldsymbol{A}_k)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) \\ \geq&L^4c_B^3n^{-1}\cdot c_B^2L^2\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right), \end{aligned}$$
where we used Lemma 101 in the penultimate line and the definition of the event $\mathscr{B}_k(c_B)$ from Section 3.3 in the last transition.
When it comes to the second term, we once again (as in the proof of Lemma 108) are going to use the fact that
$$\boldsymbol{P}_{0:k}^\perp\boldsymbol{A}_k\boldsymbol{A}^{-1} = \boldsymbol{P}_{0:k}^\perp.$$
Thus, we write
$$\begin{aligned} &\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}) \\ \geq&\mu_1(\boldsymbol{A}_k)^{-2}\mathrm{tr}(\boldsymbol{A}_k\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{A}_k) \\ \geq&\mu_1(\boldsymbol{A}_k)^{-2}\mathrm{tr}(\boldsymbol{P}_{0:k}^\perp\boldsymbol{A}_k\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{A}_k\boldsymbol{P}_{0:k}^\perp) \\ =&\mu_1(\boldsymbol{A}_k)^{-2}\mathrm{tr}(\boldsymbol{P}_{0:k}^\perp\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{P}_{0:k}^\perp) \\ =&\mu_1(\boldsymbol{A}_k)^{-2}\sum_{i>k}\lambda_i^2\boldsymbol{z}_i^\top\boldsymbol{P}_{0:k}^\perp\boldsymbol{z}_i, \end{aligned}$$

where $\boldsymbol{z}_i$ are columns of $\boldsymbol{Z}$. Note that for every fixed $i > k$ the vector $\boldsymbol{z}_i$ has i.i.d. components with variance 1 and sub-Gaussian constant at most σ_x which are independent of $\boldsymbol{P}_{0:k}$. As in the proof of Lemma 108, by Hanson-Wright inequality (Lemma 92) for some absolute constant c_1 and any $s > 0$ with probability at least $1 - 2e^{-s/c_1}$

$$\begin{aligned}|\boldsymbol{z}_i^\top \boldsymbol{P}_{0:k}^\perp \boldsymbol{z}_i - \operatorname{tr}(\boldsymbol{P}_{0:k}^\perp)| =& |\boldsymbol{z}_i^\top \boldsymbol{P}_{0:k}^\perp \boldsymbol{z}_i - (n-k)| \\ <& \sigma_x^2 \max(\sqrt{s}\|\boldsymbol{P}_{0:k}^\perp\|_F, s\|\boldsymbol{P}_{0:k}^\perp\|) \\ =& \sigma_x^2 \max(\sqrt{s}\sqrt{n-k}, s).\end{aligned}$$

So, for a large enough constant c_2 that only depends on σ_x, given that $c > c_2$ (i.e., $k < n/c_2$) for any separate i with probability at least $1 - c_2 e^{-n/c_2}$

$$\boldsymbol{z}_i^\top \boldsymbol{P}_{0:k}^\perp \boldsymbol{z}_i \geq n/c_2.$$

By Lemma 9 from [3], we can combine separate high-probability lower bounds on non-negative terms into a high-probability lower bound on the sum, that is, with probability at least $1 - 2c_2 e^{-n/c_2}$

$$\sum_{i>k} \lambda_i^2 \boldsymbol{z}_i^\top \boldsymbol{P}_{0:k}^\perp \boldsymbol{z}_i \geq \frac{n}{2c_2} \sum_{i>k} \lambda_i^2.$$

Take this event as $\mathscr{C}$.

Overall, we get that on $\mathscr{A}_k(L) \cap \mathscr{C}$

$$\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}) \geq \frac{1}{2L^2 c_2}\Lambda^{-2} n \sum_{i>k} \lambda_i^2.$$

2. Using Lemma 100 we have

$$\begin{aligned}&\|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}^2 \\ =& \left\|\boldsymbol{\Sigma}_{0:k}^{1/2}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k}\right\|^2 \\ =& \left\|\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\ \geq& \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-2}\mu_n(\boldsymbol{A}_k)^2 \left\|\left(\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1}\mu_1(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\ \geq& L^{-2}c_B^{-2}\Lambda^2 n^{-2} \cdot c_B^{-2}L^{-2}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2,\end{aligned}$$

where we used Lemma 101 in the penultimate line and the definition of the event $\mathscr{B}_k(c_B)$ from Section 3.3 in the last transition.

□

Finally, we can put everything together in the following

Lemma 109. *Suppose that the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. Take some $L > 1$. There is an absolute constant c_y and a constant c that only depends on L and σ_x, such that if $k < n/c$ and*

$$\Lambda > c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right), \tag{B.15}$$

then there exists an event $\mathscr{C}$ which only depends on $\boldsymbol{Q}$, whose probability is at least $1 - ce^{-n/c}$ such that conditionally on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B) \cap \mathscr{C}$ with probability at least $c_y^{-1} - c_y e^{-n/c}$ over the draw of $\boldsymbol{y}$ with probability at least 0.25 over the draw of $(\varepsilon_y, \varepsilon_q)$

$$\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{ridge}\|_{\boldsymbol{\Sigma}}^2 \geq c^{-1}(V + n\diamondsuit^2).$$

Proof. Take the event $\mathscr{C}$ to be the same as in Lemma 108, and denote the constant from it as c_1. By that lemma on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B) \cap \mathscr{C}$

$$\begin{aligned}\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) \geq& c_1^{-1}V,\\ \|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}^2 \geq& c_1^{-1}\Lambda^2 n^{-2}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2.\end{aligned}$$

Moreover, denote the constant from Lemma 102 as c_2. Note that $k < n/c_2$ and $\Lambda > c_2 n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}$ on if c is large enough. Thus, we by Lemma 102 on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$

$$\begin{aligned}\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\| \leq& c_2\diamondsuit,\\ \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) \leq& c_2 V\\ \|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\| \leq& c_2\Delta V.\end{aligned}$$

Now denote the constant from Lemma 107 as c_y. Combining that lemma with the results we stated above we get that conditionally on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B) \cap \mathscr{C}$ for any $t \in (0, \sqrt{n}/c_y)$ with probability at least $1 - c_y e^{-t^2/c_y}$ all the following hold almost surely over the draw of ε_q:

$$\begin{aligned}\left\|\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 \geq& c_y^{-1}c_1^{-1}V,\\ \boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y} \geq& (n-k)\mu_1(\boldsymbol{A}_k)^{-1} - c_y(t\sqrt{n} + t^2)\|\boldsymbol{A}^{-1}\|\\ \geq& n\Lambda^{-1}(L^{-1}(1 - k/n) - c_y L\Lambda^{-1}(t\sqrt{n} + t^2),\\ \boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y} \leq& nL\Lambda^{-1},\\ \|\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \leq& c_y c_2(V + t^2\Delta V),\\ |\bar{\boldsymbol{\nu}}^\top\boldsymbol{A}^{-1}\boldsymbol{y}| \leq& c_y c_2 t\diamondsuit.\end{aligned}$$

For the second inequality let's restrict t to the range $(0, \sqrt{n}/c_3)$, where c_3 is a large enough constant depending on σ_x, L, so that inequality implies $\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y} \geq c_3^{-1}n\Lambda^{-1}$.

Moreover, on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$ we can write

$$\begin{aligned}
&\left\|\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\\
\leq &\left\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\right\| \mu_n(\boldsymbol{A})^{-2} \left\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|^2\\
\leq &c_B \left(\sum_{i>k} \lambda_i^2 + n\lambda_{k+1}^2\right) \cdot L^2\Lambda^{-2} \cdot c_B n \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\\
\leq &\frac{c_B^2 L^2}{c^2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2,
\end{aligned}$$

where in the last line we used the assumption from Equation (B.15).

Combining all that with Lemma 106 gives that conditionally on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B) \cap \mathscr{C}$ for any $t \in (0, \sqrt{n}/c_3)$ with probability at least $c_y^{-1} - c_y e^{-t^2/c_y}$ over the draw of $\boldsymbol{y}$ with probability at least 0.25 over the draw of $(\varepsilon_y, \varepsilon_q)$

$$\begin{aligned}
\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}^2 \geq &\frac{1}{2}\left\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2\\
&+\frac{1}{2}(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 \left(\left\|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + 0.5\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\right)\\
&-\frac{7}{2}(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 \left\|\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\\
&-7(\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2\\
\geq &\frac{1}{2c_1 c_y} V\\
&+\frac{1}{2c_3^2} n^2 \Lambda^{-2} \cdot \left(c_1^{-1}\Lambda^2 n^{-2}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + 0.5\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\right)\\
&-\frac{7}{2} n^2\Lambda^{-2} \cdot \frac{c_B^2 L^2}{c^2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 - 7c_y^2 c_2^2 t^2 \Diamond^2 \cdot c_y c_2 (V + t^2 \Delta V).
\end{aligned}$$

If c is large enough, namely if $7c_B^2 L^2 c_3^2 < 0.25c^2$, then we have

$$\begin{aligned}
&\frac{1}{2c_3^2} n^2 \Lambda^{-2} \cdot \left(c_1^{-1}\Lambda^2 n^{-2}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + 0.5\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\right)\\
&-\frac{7}{2} n^2\Lambda^{-2} \cdot \frac{c_B^2 L^2}{c^2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\\
\geq &\frac{1}{2c_3^2} \cdot \left(c_1^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + 0.25 n^2\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\right)\\
\geq &\frac{\min(c_1^{-1}, 0.25)}{2c_3^2} n\Diamond^2/2.
\end{aligned}$$

That is, for a large enough constant c_4 that only depends on σ_x, c_B, L, we have

$$\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}^2 \geq c_4^{-1}(V + n\Diamond^2) - c_4 t^2 \Diamond^2 (V + t^2 \Delta V).$$

By Lemma 41 $V \le 2$ and $t^2 \Delta V \le 3t^2/n \le 3$. Thus,

$$\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}^2 \ge c_4^{-1}(V + n\Diamond^2(1 - 5c_4^2 t^2/n)) \ge c_5^{-1}(V + n\Diamond^2),$$

provided that c_5 is a large enough constant depending on c_4, and $t = \sqrt{n/c_5}$.

□

The ratio

Finally we can put the bound on the numerator together with the bound on the denominator and obtain the following

Theorem 43 (Main upper bound)**.** *Suppose that $\eta = 0$ — there is no label-flipping noise, and the rows of $\boldsymbol{Z}$ are σ_x-sub-Gaussian. For any $L > 1$ there are large constants a, c that only depend on σ_x and L and an absolute constant c_y such that the following holds. Suppose that $k < n/c$ and*

$$\Lambda > c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

Assume that $\boldsymbol{Q}_{k:\infty}$ is independent from $\boldsymbol{Q}_{0:k}$, and the distribution of $\boldsymbol{Q}_{k:\infty}$ is symmetric.

1. *If $N < a^{-1}\Diamond$ then with probability at least $c_y^{-1}(\mathbb{P}(\mathscr{A}_k(L)) - ce^{-n/c})_+$,*

   $$\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} < 0.$$

   Here u_+ denotes $u \vee 0$ for any $u \in \mathbb{R}$.

2. *If $N \ge a^{-1}\Diamond$ then for any $t \in (0, \sqrt{n}/c_y)$ the probability of the event*

   $$\left\{\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} \le c(1+t)\frac{N}{\sqrt{V + n\Diamond^2}}\right\}$$

   is a least

   $$(c_y^{-1} - c_y e^{-t^2/c_y} - c_y e^{-n/c})_+(\mathbb{P}(\mathscr{A}_k(L)) - ce^{-n/c})_+.$$

Proof. First of all, it is enough to show the statement for $\bar{\boldsymbol{w}}_{\text{ridge}}$ (as defined in the beginning of Section B.8) instead of $\boldsymbol{w}_{\text{ridge}}$ as it has the same distribution.

The straightforward combination of Lemmas 109 and 105 almost does the job, but we also need to show that $\bar{S} > 0$ with high probability.

Recall that

$$\bar{S} = (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})^2 + \boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\measuredangle}\bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}.$$

By Lemma 102, if c is large enough depending on σ_x, L, then on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$ we have

$$\boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\measuredangle} \ge M/c > 0.$$

Moreover, $(1+\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})^2 \geq 0$ almost surely, and $\bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}} > 0$ on $\mathscr{A}_k(L)$. Thus, if c is large enough, then $\bar{S} > 0$ on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$.

Recall that since the data is sub-Gaussian, we can take c_B in the definition of the event $\mathscr{B}_k(c_B)$ large enough depending only on σ_x such that $\mathbb{P}(\mathscr{B}_k(c_B)) \geq 1 - c_B e^{-n/c_B}$ (as shown in Section 3.3). Now the first part of the theorem is a direct consequence of part 2 of Lemma 105, while the second part of the theorem is a direct combination of Lemma 109 with the first part of 105. □

Theorem 45 (Tightness of the bounds)**.** *Suppose that the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. Suppose that $\eta = 0$ — there is no label-flipping noise. For any $L > 1$ there exist constants a, c that only depend on L, σ_x and absolute constants ε, δ such that the following holds. Suppose that $n > c$, $k < n/c$,*

$$\Lambda > c\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right),$$

and the probability of the event $\mathscr{A}_k(L)$ is at least $1-\delta$. Assume that $\boldsymbol{Q}_{k:\infty}$ is independent from $\boldsymbol{Q}_{0:k}$, and the distribution of $\boldsymbol{Q}_{k:\infty}$ is symmetric.

Then

$$\alpha_\varepsilon \leq c\frac{N}{\sqrt{V}+\sqrt{n}\diamondsuit}.$$

If additionally $N \geq a\diamondsuit$, then

$$\alpha_\varepsilon \geq c^{-1}\frac{N}{\sqrt{V}+\sqrt{n}\diamondsuit}.$$

Proof. First of all, denote the constants from Theorem 43 as $a_u, c_u, c_{y,u}$ (here index u stands for "upper bound"). Note that by that theorem, regardless whether $n\Lambda^{-1}M < a^{-1}\diamondsuit$ or $n\Lambda^{-1}M \geq a^{-1}\diamondsuit$ it still holds for any $t_u \in (0, \sqrt{n}/c_{y,u})$ that the probability of the event

$$\left\{\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}}{\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}} \leq c_u(1+t_u)\frac{n\Lambda^{-1}M}{\sqrt{V+n\diamondsuit^2}}\right\}$$

is a least

$$(c_{y,u}^{-1} - c_{y_u}e^{-t_u^2/c_{y,u}} - c_{y,u}e^{-n/c_u})_+(\mathbb{P}(\mathscr{A}_k(L)) - c_u e^{-n/c_u})_+.$$

Thus, if t_u, n, δ and ε are such that

$$(c_{y,u}^{-1} - c_{y,u}e^{-t_u^2/c_{y,u}} - c_y e^{-n/c_u})_+(1-\delta - c_u e^{-n/c_u})_+ > \varepsilon,$$

then

$$\alpha_\varepsilon < c_u(1+t_u)\frac{n\Lambda^{-1}M}{\sqrt{V+n\diamondsuit^2}}.$$

When it comes to the lower bound, recall that by Lemma 39, the event $\mathscr{B}_k(c_B)$ holds with probability at least $1 - c_B e^{-n/c_B}$ for a constant c_B that only depends on σ_x. Thus, Theorem 42 is applicable. Denote the constant from Theorem 42 as c_ℓ (here index ℓ stands for "lower bound"). Then we have that for the case $\eta = 0$ (i.e. $\boldsymbol{y} = \hat{\boldsymbol{y}}$) for any $t_\ell \in (0, \sqrt{n}/c_\ell)$, conditionally on the event $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$, with probability at least $1 - c_\ell e^{-t_\ell^2/2}$ over the draw of $\boldsymbol{y}$

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}}{\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}} \geq c_\ell^{-2} \frac{N - c_\ell^2 t_\ell \diamondsuit}{\sqrt{V + t_\ell^2 \Delta V} + \diamondsuit\sqrt{n}} \geq c_\ell^{-2} \frac{N/2}{\sqrt{V(1 + 4t_\ell^2)} + \diamondsuit\sqrt{n}},$$

where the last transition is made under the assumption that $c_\ell^2 t_\ell \diamondsuit \leq N$, and also uses the fact that $\Delta V \leq 4V$ (Lemma 41).

Thus, if we take $a = 2c_\ell^2 t_\ell$ and $t_\ell, n, \varepsilon, \delta$ are such that

$$(1 - \delta - c_B e^{-n/c_B})_+ (1 - c_\ell e^{-t_\ell^2/2})_+ \geq 1 - \varepsilon,$$

then under the condition $n\Lambda^{-1}M \geq a\diamondsuit$ we get

$$\alpha_\varepsilon \geq c_\ell^{-2} \frac{n\Lambda^{-1}M/2}{\sqrt{V(1 + 4t_\ell^2)} + \diamondsuit\sqrt{n}}.$$

Finally, to finish the proof we just need to choose c_1, t_l and t_u that can only depend on L, σ_x and absolute constants δ, ε such that for any $n > c_1$

$$\begin{aligned}
(1 - \delta - c_B e^{-n/c_B})_+ (1 - c_{y,\ell} e^{-t_\ell^2/2})_+ >& 1 - \varepsilon, \\
(c_{y,u}^{-1} - c_{y,u} e^{-t_u^2/c_{y,u}} - c_y e^{-n/c_u})_+ (1 - \delta - c_u e^{-n/c_u})_+ >& \varepsilon.
\end{aligned}$$

This is easy to do: first choose t_u large enough so that $c_{y_u} e^{-t_u^2/c_{y,u}} < c_{y_u}^{-1}/2$. Note that t_u is an absolute constant. Second, take $\varepsilon = 0.5 \wedge (c_{y_u}^{-1}/16)$ — an absolute constant. Third, take c_1 large enough depending on $c_B, c_{y,u}, c_u, \varepsilon$ so that

$$c_y e^{-c_1/c_u} \leq c_{y,u}^{-1}/4, \quad c_y e^{-c_1/c_u} \leq \frac{1}{4}, \quad c_B e^{-c_1/c_B} \leq \varepsilon/4.$$

Fourth, take $\delta = \varepsilon/4$ – an absolute constant. Finally, take t_ℓ such that $c_\ell e^{-t_\ell^2/2} \leq \varepsilon/2$ — a constant that only depends on c_ℓ (which, in its turn, only depends on L and σ_x). Combining all gives

$$\begin{aligned}
&(1 - \delta - c_B e^{-n/c_B})_+ (1 - c_{y,\ell} e^{-t_\ell^2/2})_+ \\
\geq& (1 - \varepsilon/4 - \varepsilon/4)_+ (1 - \varepsilon/2)_+ \\
>& 1 - \varepsilon, \\
&(c_{y,u}^{-1} - c_{y,u} e^{-t_u^2/c_{y,u}} - c_y e^{-n/c_u})_+ (1 - \delta - c_u e^{-n/c_u})_+ \\
>& (c_{y,u}^{-1} - c_{y,u}^{-1}/2 - c_{y,u}^{-1}/4)(1 - 1/4 - 1/4) \\
=& c_{y_u}^{-1}/16 \geq \varepsilon,
\end{aligned}$$

which finishes the proof. □

B.9 Analysis of ridge regularization

Lemma 49. *Consider a non-zero vector $\boldsymbol{v} \in \mathbb{R}^p$ and a PD symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{p \times p}$. Introduce the function $f : \mathbb{R}^p \to \mathbb{R}$ as $f(\boldsymbol{w}) = \boldsymbol{v}^\top \boldsymbol{w}/\|\boldsymbol{w}\|$. Then $f\left((\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{v}\right)$ is a non-increasing function of t on $[0, +\infty)$.*

Proof. The idea is to introduce the vector-valued function $\boldsymbol{w}(t) := (\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{v}$ and to compute the derivative of $f(\boldsymbol{w}(t))$ in t using the chain rule as

$$\frac{d}{dt} f(\boldsymbol{w}(t)) = \left(\frac{d}{dt}\boldsymbol{w}(t)\right)^\top \left(\nabla f(\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{w}(t)}\right).$$

We write the following using the brackets $\langle \boldsymbol{u}_1, \boldsymbol{u}_2 \rangle$ to denote the scalar product $\boldsymbol{u}_1^\top \boldsymbol{u}_2$.

$$\begin{aligned}
f(\boldsymbol{w}) :=& \boldsymbol{v}^\top \boldsymbol{w}/\|\boldsymbol{w}\|,\\
\nabla_{\boldsymbol{w}} f(\boldsymbol{w}) =& \boldsymbol{v}/\|\boldsymbol{w}\| - \boldsymbol{w} \cdot \boldsymbol{v}^\top \boldsymbol{w}/\|\boldsymbol{w}\|^3,\\
\dot{\boldsymbol{w}} := \frac{d}{dt}\boldsymbol{w} =& -\boldsymbol{M}(\boldsymbol{I}_p + t\boldsymbol{M})^{-2}\boldsymbol{v}\\
=& -t^{-1}(\boldsymbol{I}_p + t\boldsymbol{M} - \boldsymbol{I}_p)(\boldsymbol{I}_p + t\boldsymbol{M})^{-2}\boldsymbol{v}\\
=& -t^{-1}(\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{v} + t^{-1}(\boldsymbol{I}_p + t\boldsymbol{M})^{-2}\boldsymbol{v}\\
=& -t^{-1}\left(\boldsymbol{w} + (\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{w}\right).\\
t\boldsymbol{v}^\top \dot{\boldsymbol{w}} =& -\boldsymbol{v}^\top \boldsymbol{w} + \|\boldsymbol{w}\|^2,\\
t\boldsymbol{w}^\top \dot{\boldsymbol{w}} =& -\|\boldsymbol{w}\|^2 + \boldsymbol{w}^\top(\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{w},\\
t\|\boldsymbol{w}\|^3\langle \nabla f(\boldsymbol{w}), \dot{\boldsymbol{w}}\rangle =& \|\boldsymbol{w}\|^4 - \boldsymbol{v}^\top \boldsymbol{w} \cdot \boldsymbol{w}^\top (\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{w}\\
=& \left\langle (\boldsymbol{I}_p + t\boldsymbol{M})^{-1/2}\boldsymbol{v}, (\boldsymbol{I}_p + t\boldsymbol{M})^{-3/2}\boldsymbol{v}\right\rangle^2\\
& - \|(\boldsymbol{I}_p + t\boldsymbol{M})^{-1/2}\boldsymbol{v}\|^2 \cdot \|(\boldsymbol{I}_p + t\boldsymbol{M})^{-3/2}\boldsymbol{v}\|^2\\
\leq& 0,
\end{aligned}$$

where the last transition is by Cauchy-Schwartz . We see that the derivative of $f(\boldsymbol{w}(t))$ is non-positive when $t > 0$, thus the function $f(\boldsymbol{w}(t))$ is non-increasing in t on $[0, +\infty)$. □

Lemma 50 (Increasing the regularization cannot make the bound large)**.** *Suppose that $k < n$ and $\Lambda(\lambda) > n\lambda_k$. Then for some absolute constant $c > 0$ and any $\lambda' > \lambda$*

$$\frac{N(\lambda')}{\sqrt{V(\lambda')} + \sqrt{n}\Diamond(\lambda')} \leq c\left(1 + \frac{N(\lambda)}{\sqrt{V(\lambda)} + \sqrt{n}\Diamond(\lambda)}\right).$$

Proof. First of all, note that $\Lambda(\lambda') > \Lambda(\lambda) > n\lambda_k$. Thus, by Lemma 48 we can show that for some absolute constant $c_1 > 0$

$$\frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)} \vee \sqrt{n}\Diamond_a(\Lambda_1)} \leq c_1\left(1 + \frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)} \vee \sqrt{n}\Diamond_a(\Lambda_0)}\right),$$

where we denoted $\Lambda_0 = \Lambda(\lambda)$, $\Lambda_1 = \Lambda(\lambda')$ and used the notation from Lemma 48.

From now on we forget about the notion of k and only study the following quantity as a function of Λ:

$$\frac{N_a(\Lambda)}{\sqrt{V_a(\Lambda)}} \wedge \frac{N_a(\Lambda)}{\sqrt{n}\lozenge_a(\Lambda)}.$$

Note that if we denote

$$t := \Lambda/n, \quad \boldsymbol{v} := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}, \quad \boldsymbol{w} := (\boldsymbol{\Sigma} + t\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mu} = (\boldsymbol{I}_p + t\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{v},$$

then it becomes

$$\frac{N_a(\Lambda)}{\sqrt{n}\lozenge_a(\Lambda)} = \frac{\boldsymbol{v}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|}.$$

Thus, by Lemma 49, $\frac{N_a(\Lambda)}{\sqrt{n}\lozenge_a(\Lambda)}$ is a non-increasing function of Λ, i.e., the benefit of regularization could only potentially come from the term $\frac{N_a(\Lambda)}{\sqrt{V(\Lambda)}}$. More precisely, suppose that

$$\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\lozenge_a(\Lambda_0)} < \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\lozenge_a(\Lambda_1)}.$$

Then $\sqrt{V_a(\Lambda_0)} \geq \sqrt{n}\lozenge_a(\Lambda_0)$, otherwise we would have

$$\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\lozenge_a(\Lambda_0)} = \frac{N_a(\Lambda_0)}{\sqrt{n}\lozenge_a(\Lambda_0)} \geq \frac{N_a(\Lambda_1)}{\sqrt{n}\lozenge_a(\Lambda_1)} \geq \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\lozenge_a(\Lambda_1)}.$$

Moreover, if $\sqrt{V_a(\Lambda_1)} < \sqrt{n}\lozenge_a(\Lambda_1)$ then by Intermediate Value Theorem we can take such $\Lambda_{0.5}$ that $\sqrt{V_a(\Lambda_{0.5})} = \sqrt{n}\lozenge_a(\Lambda_{0.5})$, and we'll once again have

$$\frac{N_a(\Lambda_{0.5})}{\sqrt{V_a(\Lambda_{0.5})}} \wedge \frac{N_a(\Lambda_{0.5})}{\sqrt{n}\lozenge_a(\Lambda_{0.5})} = \frac{N_a(\Lambda_{0.5})}{\sqrt{n}\lozenge_a(\Lambda_{0.5})} \geq \frac{N_a(\Lambda_1)}{\sqrt{n}\lozenge_a(\Lambda_1)} \geq \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\lozenge_a(\Lambda_1)}.$$

In case $\Lambda_{0.5}$ as above exists set $\Lambda = \Lambda_{0.5}$, otherwise set $\Lambda = \Lambda_1$. Now we have

$$\begin{gathered}
\Lambda_1 \geq \Lambda > \Lambda_0, \\
\sqrt{V_a(\Lambda)} \geq \sqrt{n}\lozenge_a(\Lambda), \quad \sqrt{V_a(\Lambda_0)} \geq \sqrt{n}\lozenge_a(\Lambda_0), \\
\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\lozenge_a(\Lambda_0)} < \frac{N_a(\Lambda)}{\sqrt{V_a(\Lambda)}} \wedge \frac{N_a(\Lambda)}{\sqrt{n}\lozenge_a(\Lambda)} \geq \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\lozenge_a(\Lambda_1)}.
\end{gathered}$$

Let's study $\frac{N_a(\Lambda)}{\sqrt{V_a(\Lambda)}}$. The idea is to re-introduce k, but the "right one" (basically choose $k = k^*$). Then split into the $0:k$ and $k:\infty$ part and say that increasing regularization does nothing to the tail, but also cannot make the $0:k$ part more than a constant. Formally, take $k_0 = \min\{\kappa : \lambda_{\kappa+1} < \Lambda_0/n\}$. Such $k_0 < n$ exists since $\Lambda_0 > n\lambda_{k+1}$.

Now we write

$$\begin{aligned}\frac{N_a(\Lambda)}{\sqrt{V_a(\Lambda)}} =&\frac{\sum_i \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}}\\
=&\frac{\sum_{i=1}^{k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} + \frac{\sum_{i>k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}}\\
\le&\frac{\sum_{i=1}^{k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} + \frac{\sum_{i>k_0} \frac{\mu_i^2}{\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}}\\
=&\frac{\sum_{i=1}^{k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} + \frac{\sum_{i>k_0} \mu_i^2}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(n\lambda_i/\Lambda+1)^2}}}\end{aligned}$$

The second term is a decreasing function of Λ, which implies

$$\frac{\sum_{i>k_0} \mu_i^2}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(n\lambda_i/\Lambda+1)^2}}} \le \frac{\sum_{i>k_0} \mu_i^2}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(n\lambda_i/\Lambda_0+1)^2}}} = \frac{\sum_{i>k_0} \frac{\mu_i^2}{\Lambda_0/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda_0/n)^2}}} \le 2\frac{\sum_i \frac{\mu_i^2}{\lambda_i+\Lambda_0/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda_0/n)^2}}} = 2\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}},$$

where we used that $(\Lambda_0/n)^{-1} \le 2(\lambda_i + \Lambda_0/n)^{-1}$ for $i > k_0$ in the last inequality.

Now let's study the part that comes from the first k_0 components. We can write

$$\frac{\sum_{i=1}^{k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} = \frac{\sum_{i=1}^{k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_{i=1}^{k_0} \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2}}} \cdot \frac{\sqrt{\sum_{i=1}^{k_0} \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2}}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}}$$

By Lemma 49, the first multiplier is a non-increasing function of Λ. Moreover, if we plug Λ_0 instead of Λ we get

$$\begin{aligned}\frac{\sum_{i=1}^{k_0} \frac{\mu_i^2}{\lambda_i+\Lambda_0/n}}{\sqrt{\sum_{i=1}^{k_0} \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda_0/n)^2}}} \le 2\frac{\sum_{i=1}^{k_0} \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda_0/n)^2}}{\sqrt{\sum_{i=1}^{k_0} \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda_0/n)^2}}} = 2\sqrt{\sum_{i=1}^{k_0}\frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda_0/n)^2}} \le\\
\le 2\sqrt{n}\Diamond_a(\Lambda_0) \le 2\sqrt{V_a(\Lambda_0)} \le 2\sqrt{2},\end{aligned}$$

where we used that $\lambda_i \ge \Lambda_0/n$ for $i \le k_0$. In the last transition we also used that $V_a(\Lambda_0) < V(\Lambda_0)$ (Lemma 48) and $V(\Lambda_0) < 2$ (Lemma 41).

Thus, the first multiplier starts less than a constant and stays less than a constant. For the second multiplier we have

$$\frac{\sqrt{\sum_{i=1}^{k_0} \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2}}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} \le \frac{\sqrt{n}\Diamond_a(\Lambda)}{\sqrt{V_a(\Lambda)}} \le 1.$$

Overall, we've got that either

$$\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\Diamond_a(\Lambda_0)} \geq \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\Diamond_a(\Lambda_1)},$$

or

$$\frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\Diamond_a(\Lambda_1)} \leq \frac{N_a(\Lambda)}{\sqrt{V_a(\Lambda)}} \leq 2\sqrt{2} + 2\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} = 2\sqrt{2} + 2\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\Diamond_a(\Lambda_0)},$$

which implies the desired result.

□

Corollary 52 (Regularization doesn't matter for certain $\boldsymbol{\mu}$)**.** *Suppose that the distribution of the rows of $\boldsymbol{Z}$ is σ_x-sub-Gaussian. For any $L > 1$ there exist constants a, c that only depend on L, σ_x and absolute constants ε, δ such that the following holds. Suppose that $n > c$, $k < n/c$, $\mathbb{P}(\mathscr{A}_k(L, \lambda)) > 1 - \delta$, $N(\lambda) \geq a\Diamond(\lambda)$, and*

$$\Lambda(\lambda) > c\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

Suppose that $\boldsymbol{Q}_{k:\infty}$ has a symmetric distribution and is independent from $\boldsymbol{Q}_{0:k}$.

If either for some $i \leq k$

$$\boldsymbol{\mu} = \mu_i\boldsymbol{e}_i, \quad \textit{and} \quad \frac{n\lambda_i\mu_i^2}{(1 + n\lambda_i/\Lambda(\lambda))^2} \geq \sum_i \lambda_i^2,$$

(here $\boldsymbol{e}_i$ is the i-th eigenvector of $\boldsymbol{\Sigma}$), or

$$\|\boldsymbol{\mu}_{0:k}\| = 0 \quad \textit{and} \quad \sum_i \lambda_i^2 \leq n\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}},$$

then for any $\lambda' \geq \lambda$,

$$\alpha_\varepsilon(\lambda')/c \leq \alpha_\varepsilon(\lambda) \leq c\alpha_\varepsilon(\lambda').$$

Proof. Denote the constants from Theorem 45 as a_0, c_0, δ_0 and ε_0.

First of all, let's show that if a is chosen to be equal to a_0, then for any $\lambda' \geq \lambda$ it holds $n\Lambda(\lambda')^{-1}M(\lambda') \geq a_0\Diamond(\lambda')$. Indeed, if $\|\boldsymbol{\mu}_{0:k}\| = 0$, then

$$\begin{aligned}
&\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\Diamond(\lambda')} \\
=&\frac{\left\|\left(\Lambda(\lambda')n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2}{\sqrt{n^{-1}\left\|\left(\Lambda(\lambda')n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}}} \\
=&\frac{\sqrt{n}\|\boldsymbol{\mu}_{k:\infty}\|^2}{\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}}
\end{aligned}$$

— doesn't depend on λ' at all.

In the case that $\boldsymbol{\mu}$ is an eigenvector of $\boldsymbol{\Sigma}$, that is, $\boldsymbol{\mu} = \mu_i \boldsymbol{e}_i$ for some $i \in [k]$, we have

$$\begin{aligned}&\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\diamondsuit(\lambda')}\\ =&\frac{\left(1+\lambda_i^{-1}n^{-1}\Lambda(\lambda')\right)^{-1}\lambda_i^{-1}\mu_i^2}{\sqrt{n^{-1}\left(1+\lambda_i^{-1}n^{-1}\Lambda(\lambda')\right)^{-1}\lambda_i^{-1}\mu_i^2}}\\ =&\sqrt{n/\lambda_i}\end{aligned}$$

— doesn't depend on λ' once again.

That is

$$\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\diamondsuit(\lambda')} = \frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\diamondsuit(\lambda)} \geq a = a_0.$$

Note also that

$$\Lambda(\lambda') \geq \Lambda(\lambda) > c\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right),$$

and

$$\mathbb{P}(\mathscr{A}_k(L,\lambda')) \geq \mathbb{P}(\mathscr{A}_k(L,\lambda)) \geq 1-\delta_0,$$

so if $c > c_0$, $\varepsilon = \varepsilon_0$ and $\delta = \delta_0$ then the assumptions of Theorem 45 are satisfied for both λ and λ', which yields

$$\alpha_\varepsilon(\lambda)/c_0 \leq \frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{V(\lambda)}+\sqrt{n}\diamondsuit(\lambda)} \leq c_0\alpha_\varepsilon(\lambda), \quad \alpha_\varepsilon(\lambda')/c_0 \leq \frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{V(\lambda')}+\sqrt{n}\diamondsuit(\lambda')} \leq c_0\alpha_\varepsilon(\lambda').$$

We've already seen that $\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\diamondsuit(\lambda')} = \frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{n}\diamondsuit(\lambda)}$, so the only thing we need to study is $V(\lambda)$. Namely, we are going to show that under the assumptions we made $V(\lambda) < n\diamondsuit^2(\lambda)$ and $V(\lambda') < n\diamondsuit^2(\lambda')$. That will finish the proof since in that case

$$\begin{aligned}\alpha_\varepsilon(\lambda')/c_0 \leq \frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{V(\lambda')}+\sqrt{n}\diamondsuit(\lambda')} \leq \frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\diamondsuit(\lambda')} =&\\ = \frac{n\Lambda(\lambda')^{-1}M(\lambda)}{\sqrt{n}\diamondsuit(\lambda)} \leq 2\frac{n\Lambda(\lambda')^{-1}M(\lambda)}{\sqrt{V(\lambda')}+\sqrt{n}\diamondsuit(\lambda)} \leq 2c_0\alpha_\varepsilon(\lambda),\end{aligned}$$

and analogously $\alpha_\varepsilon(\lambda)/c_0 \leq 2c_0\alpha_\varepsilon(\lambda')$.

Thus, in the rest of the proof we show that $V(\lambda) < n\diamondsuit^2(\lambda)$ and $V(\lambda') < n\diamondsuit^2(\lambda')$. Let's write out two cases:

1. $\boldsymbol{\mu}$ is supported on the tail. Then

$$\begin{aligned} V(\lambda) :=& n^{-1}\mathrm{tr}\left(\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda(\lambda)^{-2}n\sum_{i>k}\lambda_i^2, \\ \Diamond^2 :=& n^{-1}\left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \\ =& n\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \end{aligned}$$

We want to show that

$$n^{-1}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1} + n\Lambda(\lambda)^{-1}\boldsymbol{I}_k\right)^{-2}\right) + n^{-1}\sum_{i>k}\lambda_i^2 \le \|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}},$$

which holds since

$$\begin{aligned} & n^{-1}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1} + n\Lambda(\lambda)^{-1}\boldsymbol{I}_k\right)^{-2}\right) + n^{-1}\sum_{i>k}\lambda_i^2 \\ <& n^{-1}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1}\right)^{-2}\right) + n^{-1}\sum_{i>k}\lambda_i^2 \\ =& \sum_i \lambda_i^2 \\ \le& n\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}. \end{aligned}$$

We showed that $V(\lambda) \le n\Diamond^2(\lambda)$. Note that $V(\lambda') \le n\Diamond^2(\lambda')$ by exactly the same argument.

2. $\boldsymbol{\mu} = \boldsymbol{e}_i$ for $i \le k$. Write out V and $\Diamond$ again:

$$\begin{aligned} V(\lambda) :=& n^{-1}\mathrm{tr}\left(\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda(\lambda)^{-2}n\sum_{i>k}\lambda_i^2, \\ \Diamond(\lambda)^2 :=& n^{-1}\left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \\ =& n^{-1}\frac{\lambda_i^{-1}\mu_i^2}{(n^{-1}\Lambda(\lambda)\lambda_i^{-1} + 1)^2} \\ =& \frac{n\lambda_i\mu_i^2}{(\Lambda(\lambda) + n\lambda_i)^2} \end{aligned}$$

By the same argument as before, since

$$\frac{n\lambda_i\mu_i^2}{(1 + n\lambda_i/\Lambda(\lambda))^2} \ge \sum_i \lambda_i^2,$$

we have $V(\lambda) < n\Diamond(\lambda)^2$. When it comes to λ', we can write

$$\frac{n\lambda_i\mu_i^2}{(1 + n\lambda_i/\Lambda(\lambda'))^2} \ge \frac{n\lambda_i\mu_i^2}{(1 + n\lambda_i/\Lambda(\lambda))^2} \ge \sum_i \lambda_i^2,$$

which yields $V(\lambda') < n\Diamond(\lambda')^2$.

□

Lemma 53. *For any $\sigma_x \geq 1, L > 1$ there exist constants a, c that only depend on σ_x and absolute constants ε, δ such that the following holds. Suppose that $n > c$, $0 < k < n/c$. Take any $C > 1$ and construct the classification problem as follows:*

1. *Take $\boldsymbol{Z}_{k:\infty}$ with σ_x-sub-Gaussian rows and the sequence $\{\lambda_i\}_{i>k}$ and regularization parameter λ such that $\mathbb{P}(\mathscr{A}_k(L, \lambda)) \geq 1 - \delta$ and*
$$\Lambda(\lambda) \geq c\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$
2. *Take $\boldsymbol{Z}_{0:k}$ with σ_x-sub-Gaussian rows independent from $\boldsymbol{Z}_{k:\infty}$, and $\{\lambda_i\}_{i=1}^k$ such that $n\lambda_k \geq C\Lambda(\lambda)$.*
3. *Take $\boldsymbol{\mu}_{k:\infty}$ whose most energy is spread among the eigendirections of $\boldsymbol{\Sigma}$ with small eigenvalues, that is,*
$$\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \leq C^{-1}n^{-1}\Lambda(\lambda)\|\boldsymbol{\mu}_{k:\infty}\|^2.$$
4. *Take[1] $\boldsymbol{\mu}_{0:k}$ which balances $\boldsymbol{\mu}_{k:\infty}$ in the following sense:*
$$nC^{-1}\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2 \geq \|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2 \geq n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2. \tag{3.23}$$
5. *Scale $\boldsymbol{\mu}$ up[2] if needed, so it holds that*
$$n\Diamond^2(\lambda) \geq V(\lambda) \quad \text{and } N(\lambda) \geq a\Diamond(\lambda).$$

Then for any λ' such that $\Lambda(\lambda') \geq C\Lambda(\lambda)$
$$\alpha_\varepsilon(\lambda) \geq \frac{C}{c}\alpha_\varepsilon(\lambda').$$

Proof. Take a, ε, δ to be the same as in Theorem 45. Denote the constant c from that theorem as c_1. In the end we will take c large enough depending on c_1. If $c > c_1$ then Theorem 45 implies that
$$\alpha_\varepsilon(\lambda) \geq c_1^{-1}\frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{V(\lambda)} + \sqrt{n}\Diamond(\lambda)} \geq c_1^{-1}\frac{n\Lambda(\lambda)^{-1}M(\lambda)}{2\sqrt{n}\Diamond(\lambda)}.$$

At the same time, by Theorem 45 for any $\lambda' > \lambda$
$$\alpha_\varepsilon(\lambda') \leq c_1\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{V(\lambda')} + \sqrt{n}\Diamond(\lambda')} \leq c_1\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\Diamond(\lambda')}.$$

[1] Note that such $\boldsymbol{\mu}_{0:k}$ exists because of how we chose $\boldsymbol{\mu}_{k:\infty}$.

[2] Note that the previous conditions were homogeneous in $\boldsymbol{\mu}$, so multiplying it by a scalar does not break them.

Take such $\hat{\lambda}$ that $\Lambda(\hat{\lambda}) = C\Lambda(\lambda)$. Note that $\hat{\lambda} > \lambda$. By our construction, due to the fact that $\lambda_k > C\Lambda(\lambda) = \Lambda(\hat{\lambda})$ and Equation (3.23) we have

$$\begin{aligned}
n\Lambda(\lambda)^{-1}M(\lambda) \geq& n\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2,\\
n\Lambda(\hat{\lambda})^{-1}M(\hat{\lambda}) \leq& \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}} + n\Lambda(\hat{\lambda})^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2\\
\leq& 2n\Lambda(\hat{\lambda})^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2,\\
n\Diamond(\lambda)^2 \leq& \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}} + n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\
\leq& 2\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}},\\
n\Diamond(\hat{\lambda})^2 \geq& \left\|\left(\Lambda(\hat{\lambda})n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\
\geq& \frac{1}{4}\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}},
\end{aligned}$$

where we used that $\Lambda(\hat{\lambda})n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k$ is a diagonal matrix whose diagonal elements are at most 2 in the last transition.

Combining everything together we get that

$$\begin{aligned}
\alpha_\varepsilon(\lambda) \geq c_1^{-1}\frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{n}\Diamond(\lambda)} \geq \frac{1}{2c_1}\frac{n\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2}{\sqrt{2}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}} =\\
= \frac{C}{8\sqrt{2}c_1}\frac{2n\Lambda(\hat{\lambda})^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2}{\frac{1}{2}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}} \geq \frac{C}{8\sqrt{2}c_1^2}\cdot c_1\frac{n\Lambda(\hat{\lambda})^{-1}M(\hat{\lambda})}{\sqrt{n}\Diamond(\hat{\lambda})} \geq \frac{C}{8\sqrt{2}c_1^2}\alpha_\varepsilon(\hat{\lambda}).
\end{aligned}$$

We obtained the result for $\lambda' = \hat{\lambda}$. To extend the result for all $\lambda' > \hat{\lambda}$ note that by Lemma 48 and the fact that $N_a(\lambda)/\Diamond_a(\lambda)$ is a non-increasing function of λ for some absolute constant $c_2 > 1$ we can write

$$\begin{aligned}
\alpha_\varepsilon(\lambda) \geq \frac{C}{8\sqrt{2}c_1^2}\cdot c_1\frac{n\Lambda(\hat{\lambda})^{-1}M(\hat{\lambda})}{\sqrt{n}\Diamond(\hat{\lambda})} \geq \frac{C}{8\sqrt{2}c_1^2c_2}\cdot c_1\frac{N_a(\hat{\lambda})}{\sqrt{n}\Diamond_a(\hat{\lambda})} \geq\\
\geq \frac{C}{8\sqrt{2}c_1^2c_2}\cdot c_1\frac{N_a(\lambda')}{\sqrt{n}\Diamond_a(\lambda')} \geq \frac{C}{8\sqrt{2}c_1^2c_2^2}\cdot c_1\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\Diamond(\lambda')} \geq \frac{C}{8\sqrt{2}c_1^2c_2^2}\alpha_\varepsilon(\lambda').
\end{aligned}$$

Taking $c = 8\sqrt{2}c_1^2c_2^2$ finishes the proof. □

Corollary 54. *There exists absolute constants a, b such that the following holds. Take $p = \infty$, $n > a$ and $1 \leq k < n/a$. Consider the following classification problem with Gaussian data (in infinite dimension) and no label-flipping noise ($\eta = 0$):*

$$\lambda_i = \begin{cases} 2b, & i \leq k,\\ e^{-(i-k)/(bn)}, & i > k.\end{cases}, \quad \mu_i = \begin{cases} 4\sqrt{b/k}, & i \leq k,\\ 4\sqrt{b}\cdot 2^{-(i-k)/2}, & i > k.\end{cases}$$

Then the value of λ that maximizes $\alpha_\varepsilon(\lambda)$ is negative.

Proof. Since we consider Gaussian data, σ_x is an absolute constant. Let's take $L = 2$ and denote the corresponding constant c from Lemma 53 to be c_1. We are going to use that lemma to construct such distribution of data that the quantile $\alpha_\varepsilon(\lambda)$ is minimized for a negative λ. Note that to do that it is enough to take $C = c_1$ and $\lambda = -\frac{c_1-1}{c_1}\sum_{i>k}\lambda_i$ as this condition is equivalent to $\Lambda(0) = c_1\Lambda(\lambda)$.

Let's take infinite-dimensional Gaussian data with slow exponential decay in the tail, that is

$$\lambda_i = \begin{cases} \ell, & i \le k, \\ e^{-\alpha(i-k)}, & i > k. \end{cases}, \quad \mu_i = \begin{cases} m_{0:k}, & i \le k, \\ m_{k:\infty} e^{-\frac{\beta}{2}(i-k)}, & i > k. \end{cases}$$

Thus, the whole classification problem is described by scalars $\ell, \alpha, \beta, m_{0:k}, m_{k:\infty}, n, k$.

Let's see how we need to choose those scalars in order to follow the recipe from Lemma 53. This is an absolute constant since L is an absolute constant and the data is Gaussian. As discussed before, we fix $C = c_1$ and put $\lambda = -\frac{c_1-1}{c_1}\sum_{i>k}\lambda_i$

Due to Lemma 37, for Gaussian data the statement $\mathbb{P}(\mathscr{A}_k(L,\lambda)) \ge 1-\delta$ follows from the statement

$$\Lambda(\lambda) \ge b\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right),$$

where b is a large constant that depends on δ. Let's also take it to be larger than c_1 in order to fully satisfy step 1 of Lemma 53. For our covariance and regularization this translates into

$$\frac{1}{c_1}\frac{e^{-\alpha}}{1-e^{-\alpha}} \ge b\left(ne^{-\alpha} \vee \sqrt{n\frac{e^{-2\alpha}}{1-e^{-2\alpha}}}\right),$$

which can be equivalently rewritten as

$$1 - e^{-\alpha} \le \frac{1}{bc_1\sqrt{n}}\left(\frac{1}{\sqrt{n}} \wedge \sqrt{1-e^{-2\alpha}}\right).$$

For $x \in (0,1)$ it holds that $1 - x < e^{-x} < 1 - x(1-e^{-1}) < 1 - x/2$. Thus, assuming that $\alpha < 0.5$ the condition above follows from the following:

$$\alpha \le \frac{1}{bc_1\sqrt{n}}\left(\frac{1}{\sqrt{n}} \wedge \sqrt{\alpha}\right),$$

that is

$$\alpha \le \frac{1}{bc_1n} \wedge \frac{1}{b^2c_1^2n} = \frac{1}{b^2c_1^2n}.$$

Let's take $c_2 = b^2c_1^2$ and put $\alpha = c_2^{-1}n^{-1}$. Then conditions from step 1 of Lemma 53 are satisfied.

The second part of Lemma 53 states that we require $n\lambda_k \ge C\Lambda(\lambda)$, that is, $n\ell \ge C\left(\frac{e^{-\alpha}}{1-e^{-\alpha}} + \lambda\right) = \frac{e^{-\alpha}}{1-e^{-\alpha}}$. Note that we also need $\ell \ge 1$ in order for the sequence $\{\lambda_i\}$

to be non-increasing. Given our previous choice of α we have

$$\frac{e^{-\alpha}}{1-e^{-\alpha}} \le \frac{2}{\alpha} = 2c_2 n.$$

Thus, we can take $\ell = 2c_2$.

The third part of Lemma 53 requires the following

$$\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \le C^{-1} n^{-1} \Lambda(\lambda) \|\boldsymbol{\mu}_{k:\infty}\|^2,$$

which we equivalently transform below:

$$\begin{aligned}\sum_{i>k} e^{-(\alpha+\beta)(i-k)} &\le C^{-2} n^{-1} \frac{e^{-\alpha}}{1-e^{-\alpha}} \sum_{i>k} e^{-\beta(i-k)},\\ \frac{e^{-(\alpha+\beta)}}{1-e^{-(\alpha+\beta)}} &\le C^{-2} n^{-1} \frac{e^{-\alpha}}{1-e^{-\alpha}} \frac{e^{-\beta}}{1-e^{-\beta}},\\ \frac{1-e^{-\beta}}{1-e^{-(\alpha+\beta)}} &\le C^{-2} n^{-1} \frac{1}{1-e^{-\alpha}}.\end{aligned}$$

Let's restrict the range of β so that $\alpha + \beta < 1$. Then it is sufficient to choose β that satisfies the following stronger condition:

$$\frac{2\beta}{\alpha+\beta} \le \frac{1}{C^2 n \alpha}.$$

Plugging the expression for α yields

$$2\beta \le \frac{c_2}{C^2}\left(\beta + \frac{1}{c_2 n}\right).$$

Actually, since $c_2 = b^2 C^2 > 2C^2$, the inequality above always holds, so we can take any $\beta < 1 + \alpha$, for example, $\beta = \ln(2)$ (to make further computations simpler).

Next, part 4 of Lemma 53 requires

$$nC^{-1}\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2 \ge \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \ge n^2 \Lambda(\lambda)^{-2} \|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}},$$

that is,

$$n\frac{1-e^{-\alpha}}{e^{-\alpha}} \frac{e^{-\beta}}{1-e^{-\beta}} \ge \frac{k m_{0:k}^2}{\ell m_{k:\infty}^2} \ge n^2 C^2 \frac{(1-e^{-\alpha})^2}{e^{-2\alpha}} \frac{e^{-(\alpha+\beta)}}{1-e^{-(\alpha+\beta)}}.$$

Plugging in $\beta = \ln(2)$ and simplifying yields

$$n\frac{1-e^{-\alpha}}{e^{-\alpha}} \ge \frac{k m_{0:k}^2}{\ell m_{k:\infty}^2} \ge n^2 C^2 \frac{(1-e^{-\alpha})^2}{e^{-\alpha}} \frac{1}{2-e^{-\alpha}}.$$

As before, let's replace it by a stronger condition:

$$\frac{n\alpha}{2} \ge \frac{k m_{0:k}^2}{\ell m_{k:\infty}^2} \ge 2n^2 C^2 \alpha^2.$$

Plugging in $\alpha = (c_2 n)^{-1}$ and $\ell = 2c_2$ yields

$$\begin{aligned}\frac{1}{2c_2} \geq \frac{km_{0:k}^2}{2c_2 m_{k:\infty}^2} \geq \frac{2C^2}{c_2^2},\\ 1 \geq \frac{km_{0:k}^2}{m_{k:\infty}^2} \geq \frac{4C^2}{c_2} = \frac{4}{b^2}.\end{aligned}$$

We see that since $b > 4$, we can simply put $km_{0:k}^2/m_{k:\infty}^2 = 1$, and part 4 of Lemma 53 is satisfied.

At last, we need to check the last part of the lemma. Let's start with writing out and transforming the expressions for $n\Diamond^2(\lambda), V(\lambda)$, and $n\Lambda^{-1}(\lambda)M(\lambda)$:

$$\begin{aligned}\Lambda(\lambda) =& C^{-1}\frac{e^{-\alpha}}{1-e^{-\alpha}} \in \left(\frac{1}{2C\alpha}, \frac{2}{C\alpha}\right),\\ V(\lambda) :=& n^{-1}\mathrm{tr}\left(\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2} n \sum_{i>k}\lambda_i^2\\ =& n^{-1}\frac{k}{\left(1 + n^{-1}\ell^{-1}C^{-1}\frac{e^{-\alpha}}{1-e^{-\alpha}}\right)^2} + C^2 n e^{2\alpha}(1-e^{-\alpha})^2\frac{e^{-2\alpha}}{1-e^{-2\alpha}}\\ \leq& \frac{k/n}{\left(1 + n^{-1}\ell^{-1}C^{-1}/(2\alpha)\right)^2} + C^2 n\frac{\alpha^2}{\alpha}\\ =& \frac{k/n}{\left(1 + n^{-1}(2c_2)^{-1}C^{-1}c_2 n/2\right)^2} + C^2 n\alpha\\ =& \frac{k/n}{\left(1 + 0.25C^{-1}\right)^2} + C^2 c_2^{-1}\\ \leq& 2.\end{aligned}$$

$$\begin{aligned}n\Lambda^{-1}(\lambda)M(\lambda) :=& \left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda(\lambda)^{-1}(\lambda)\|\boldsymbol{\mu}_{k:\infty}\|^2\\ =& k\frac{m_{0:k}^2}{\ell(1 + n^{-1}\ell^{-1}\Lambda(\lambda))} + m_{k:\infty}^2 n\Lambda(\lambda)^{-1}\frac{e^{-\beta}}{1-e^{-\beta}}\\ \geq& k\frac{m_{0:k}^2}{\ell(1 + 2n^{-1}\ell^{-1}/(C\alpha))} + \frac{1}{2}m_{k:\infty}^2 nC\alpha(\lambda)\frac{e^{-\beta}}{1-e^{-\beta}}\\ =& \frac{km_{0:k}^2}{2c_2 + 2c_2/C} + \frac{1}{2}m_{k:\infty}^2 nC/(c_2 n)\\ \geq& \frac{1}{4c_2}(km_{0:k}^2 + m_{k:\infty}^2) = \frac{m_{k:\infty}^2}{2c_2}.\end{aligned}$$

$$\begin{aligned}
n\diamondsuit^2(\lambda) :=& \left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\
=&k\frac{m_{0:k}^2}{\ell(1+n^{-1}\ell^{-1}\Lambda(\lambda))^2} + m_{k:\infty}^2 n^2\Lambda(\lambda)^{-2}\frac{e^{-(\alpha+\beta)}}{1-e^{-(\alpha+\beta)}}\\
=&\frac{2c_2 m_{k:\infty}^2}{(2c_2+\Lambda(\lambda)/n)^2} + m_{k:\infty}^2 n^2\Lambda(\lambda)^{-2}\frac{e^{-\alpha}}{2-e^{-\alpha}}
\end{aligned}$$

For $n\diamondsuit^2$ we need bounds from both sides. In what follows we write them separately.

$$\begin{aligned}
n\diamondsuit^2(\lambda) \le& \frac{2c_2 m_{k:\infty}^2}{(2c_2+0.5n^{-1}C^{-1}\alpha^{-1})^2} + m_{k:\infty}^2 n^2(2C\alpha)^2\\
=&\frac{2c_2 m_{k:\infty}^2}{(2c_2+0.5c_2C^{-1})^2} + m_{k:\infty}^2(2C/c_2)^2 \le \qquad\qquad 5m_{k:\infty}^2/c_2,
\end{aligned}$$

where in the last transition we used that $c_2/C = b > \sqrt{c_2}$.

When it comes to the bound from below, we write

$$\begin{aligned}
n\diamondsuit^2(\lambda) \ge& \frac{2c_2 m_{k:\infty}^2}{(2c_2+2n^{-1}C^{-1}\alpha^{-1})^2} + m_{k:\infty}^2 n^2(2C\alpha)^2/3\\
=&\frac{2c_2 m_{k:\infty}^2}{(2c_2+2c_2C^{-1})^2} + m_{k:\infty}^2(2C/c_2)^2/3\\
\ge& m_{k:\infty}^2/(8c_2),
\end{aligned}$$

where we used $2c_2 + 2c_2C^{-1} \le 4c_2$ in the last transition.

Finally, we can write out the conditions from part 5 of Lemma 53. According to the bounds above, the following conditions on $m_{k:\infty}$ are sufficient:

$$m_{k:\infty}^2/(8c_2) \ge 2, \quad \frac{m_{k:\infty}^2}{2c_2} \ge a\sqrt{\frac{5}{c_2 n}}m_{k:\infty},$$

that is $m_{k:\infty}^2 \ge (16c_2) \vee (5c_2a^2/n) = 16c_2$ given that n is large enough. □

Lemma 55. *For any $\sigma_x > 1, L > 1$ there exist constants a, c that only depend on L, σ_x and absolute constants ε, δ such that the following holds. Suppose that $n > c$, $0 < k < n/c$. Take any $C > 1$ and construct the classification problem as follows:*

1. *Take $\boldsymbol{Z}_{k:\infty}$ with σ_x-sub-Gaussian rows and the sequence $\{\lambda_i\}_{i>k}$ and regularization parameter λ such that $\mathbb{P}(\mathscr{A}_k(L,\lambda)) \ge 1-\delta$ and*

$$\Lambda(\lambda) > c\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

2. *Take $\boldsymbol{Z}_{0:k}$ with σ_x-sub-Gaussian rows independent from $\boldsymbol{Z}_{k:\infty}$, and $\{\lambda_i\}_{i=1}^k$ such that $n\lambda_k \geq \Lambda(\lambda)$.*

3. *Take $\boldsymbol{\mu}$ that is only supported on the first k coordinates (i.e., $\|\boldsymbol{\mu}_{k:\infty}\| = 0$) such that*

$$\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}} \|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}} \geq C\|\boldsymbol{\mu}_{0:k}\|^2. \tag{3.24}$$

4. *Scale $\boldsymbol{\mu}$ up if needed, so that*

$$n\Diamond^2(\lambda) \geq V(\lambda) \quad \textit{and } N(\lambda) \geq a\Diamond(\lambda).$$

Then for any λ' such that $\Lambda(\lambda') \geq n\lambda_1$

$$\alpha_\varepsilon(\lambda) \geq \frac{C}{c}\alpha_\varepsilon(\lambda').$$

Proof. Take a, ε, δ to be the same as in Theorem 45. Denote the constant c from that theorem as c_1. In the end we will take c large enough depending on c_1. If $c > c_1$ then Theorem 45 implies that

$$\alpha_\varepsilon(\lambda) \geq c_1^{-1}\frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{V(\lambda)} + \sqrt{n}\Diamond(\lambda)} \geq c_1^{-1}\frac{n\Lambda(\lambda)^{-1}M(\lambda)}{2\sqrt{n}\Diamond(\lambda)}.$$

At the same time, by Theorem 45 for any $\lambda' > \lambda$

$$\alpha_\varepsilon(\lambda') \leq c_1\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{V(\lambda')} + \sqrt{n}\Diamond(\lambda')} \leq c_1\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\Diamond(\lambda')}.$$

Since $\|\boldsymbol{\mu}_{k:\infty}\| = 0$, $\Lambda(\lambda) \leq n\lambda_k$ and $\Lambda(\lambda') > n\lambda_1$, we can write

$$\begin{aligned}
n\Lambda(\lambda)^{-1}M(\lambda) =& \left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\
\geq& \frac{1}{2}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2,\\
n\Lambda(\lambda')^{-1}M(\lambda') =& \left\|\left(\Lambda(\lambda')n^{-1}\boldsymbol{I}_k + \boldsymbol{\Sigma}_{0:k}\right)^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\
\leq& 2n\Lambda(\lambda')^{-1}\|\boldsymbol{\mu}_{0:k}\|^2,\\
n\Diamond(\lambda)^2 \leq& \|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2,\\
n\Diamond(\lambda')^2 =& \left\|\left(\Lambda(\lambda')n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\
\geq& \frac{1}{4}n^2\Lambda(\lambda')^{-2}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}^2,
\end{aligned}$$

where we used the fact that $\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k$ and $\Lambda(\lambda')n^{-1}\boldsymbol{I}_k + \boldsymbol{\Sigma}_{0:k}$ are both diagonal matrices whose diagonal elements are at most 2.

Combining everything together we get that

$$\alpha_\varepsilon(\lambda) \geq c_1^{-1}\frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{n}\Diamond(\lambda)} \geq \frac{1}{2c_1}\frac{\frac{1}{2}\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}}}{\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}^{-1}_{0:k}}} = \frac{1}{4c_1}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}^{-1}_{0:k}} \geq \frac{C}{4c_1}\frac{\|\boldsymbol{\mu}_{0:k}\|^2}{\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}} =$$

$$= \frac{C}{16c_1}\frac{2n\Lambda(\lambda')^{-1}\|\boldsymbol{\mu}_{0:k}\|^2}{\frac{1}{2}n\Lambda(\lambda')^{-1}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}} \geq \frac{C}{16c_1^2}\cdot c_1\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\Diamond(\lambda')} \geq \frac{C}{16c_1^2}\alpha_\varepsilon(\lambda').$$

Taking $c = 16c_1^2$ finishes the proof. □

Corollary 57. *There exist absolute constants $b > c$ such that the following holds. Take $p > bn$, and $b \leq k < n/b$. Consider the following classification problem with Gaussian data (in dimension p) and no label-flipping noise ($\eta = 0$):*

$$\lambda_i = \begin{cases} k^{-4i/k}, & i \leq k, \\ \frac{cn}{pk^4}, & i > k. \end{cases}, \quad \mu_i = \begin{cases} \frac{b\ln(k)}{k^5}\left(\frac{k}{n} + \frac{n}{p}\right), & i \leq k, \\ 0, & i > k. \end{cases}$$

Then the value of λ that maximizes $\alpha_\varepsilon(\lambda)$ is negative.

Proof. Since we consider Gaussian data, σ_x is an absolute constant. Let's take $L = 2$ and denote the corresponding constants a, c from Lemma 55 to be a_1, c_1. We are going to use that lemma to construct such distribution of data that the quantile $\alpha_\varepsilon(\lambda)$ is minimized for a negative λ. Note that to do that it is enough to take $C = c_1$ and $\lambda = -\frac{c_1 - 1}{c_1}\sum_{i>k}\lambda_i$ as this condition is equivalent to $\Lambda(0) = c_1\Lambda(\lambda)$.

Let's take finite-dimensional Gaussian data with exponential decay in the first k components and isotropic tails:

$$\lambda_i = \begin{cases} e^{-\alpha i}, & i \leq k, \\ \ell, & i > k. \end{cases}$$

Note that $\Lambda(\lambda) = \Lambda(0)/c_1 = \ell(p-k)/c_1$.

We take $\boldsymbol{\mu}_{k:\infty}$ to be zero, in accordance with Lemma 55. When it comes to $\boldsymbol{\mu}_{0:k}$, we just put all its components to be equal, that is

$$\mu_i = \begin{cases} m, & i \leq k, \\ 0, & i > k. \end{cases}$$

Thus, the whole classification problem is described by scalars ℓ, m, α, n, k, p.

Our goal is to find the values of these parameters such that the conditions from Lemma 55 are satisfied for some L. We take $L = 2$.

The first part of that lemma says that $\mathscr{A}_k(2)$ should be satisfied with probability at least $1 - \delta$ and that

$$\ell(p-k)/c_1 = \Lambda(\lambda) \geq c_1\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right) = c_1\ell(n \vee \sqrt{n(p-k)}).$$

Due to Lemma 37, for Gaussian data both these conditions follow from $p > bn$ and $n > b$, where b is a large enough absolute constant.

The second part of Lemma 55 requires $n\lambda_i > \Lambda(\lambda)$, that is, $ne^{-\alpha k} \geq \ell(p-k)/c_1$, so we can take $\ell = c_1 n e^{-\alpha k}/p$. Note that since $p > c_1 n$ we have $\lambda_{k+1} = \ell < e^{-\alpha k} = \lambda_k$, so the eigenvalues remain in the right order.

The third part of Lemma 55 demands $\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}} \geq C\|\boldsymbol{\mu}_{0:k}\|^2$, that is,

$$\sqrt{\left(m^2\sum_{i=1}^k e^{-\alpha i}\right)\left(m^2\sum_{i=1}^k e^{\alpha i}\right)} \geq c_1 k m^2,$$
$$\sqrt{\frac{1-e^{-k\alpha}}{1-e^{-\alpha}}\frac{e^{k\alpha}-1}{e^{\alpha}-1}} \geq c_1 k.$$

Note that for $\alpha > 0$ we have $1 - e^{-k\alpha} > 1 - e^{-\alpha}$. Moreover, $e^{k\alpha} - 1 > (e^{\alpha} - 1)e^{(k-1)\alpha}$. Thus, it is enough to satisfy the following weaker condition:

$$e^{(k-1)\alpha/2} \geq c_1 k,$$

so we need to take $\alpha \geq 2\ln(c_1 k)/(k-1)$. Since k is lower bounded by a large constant b, we can take $\alpha = 4\ln(k)/k$. Plugging it into equation for ℓ yields $\ell = c_1 n e^{-\alpha k}/(ep) = c_1 n/(epk^4)$. We take $c = c_1/e$, so $\ell = cnp^{-1}k^{-4}$.

Finally, we need to take m large enough so that part 4 of Lemma 55 is satisfied. To check that part, we start with writing the expressions for $n\lozenge^2(\lambda), V(\lambda)$, and $n\Lambda^{-1}(\lambda)M(\lambda)$ and bounding them.

$$\begin{aligned}
\Lambda(\lambda) =&\ell(p-k)/c_1 = \frac{n(p-k)}{epk^4},\\
V(\lambda) :=&n^{-1}\mathrm{tr}\left(\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2}n\sum_{i>k}\lambda_i^2\\
\leq&\frac{k}{n} + \Lambda^{-2}n\sum_{i>k}\lambda_i^2\\
=&\frac{k}{n} + \frac{n}{p-k}.
\end{aligned}$$

$$\begin{aligned}
n\Lambda^{-1}(\lambda)M(\lambda) :=& \left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda(\lambda)^{-1}(\lambda)\|\boldsymbol{\mu}_{k:\infty}\|^2\\
=& m^2\sum_{i=1}^{k}\frac{e^{\alpha i}}{1+e^{\alpha i}(p-k)/(epk^4)}\\
\geq& \frac{m^2}{1+e^{\alpha k}/(ek^4)}\sum_{i=1}^{k}e^{\alpha i}\\
=& \frac{m^2}{1+e^{-1}}\sum_{i=1}^{k}e^{\alpha i}.
\end{aligned}$$

Let's bound the sum of exponents separately. We write

$$\sum_{i=1}^{k}e^{\alpha i} = e^{\alpha}\frac{e^{\alpha k}-1}{e^{\alpha}-1} = \frac{k^4-1}{1-e^{-4\ln(k)/k}}\begin{cases}\geq \frac{k^5}{5\ln(k)}\\ \leq \frac{k^5}{2\ln(k)}\end{cases}$$

where we used that $\alpha < 1$ (since k is large enough) and for $x \in (0,1)$ it holds $x > 1 - e^{-x} > x/2$.

Thus,

$$n\Lambda^{-1}(\lambda)M(\lambda) \geq \frac{m^2k^5}{10\ln(k)}.$$

When it comes to $\diamondsuit$, the derivation is very similar as above:

$$\begin{aligned}
n\diamondsuit^2(\lambda) :=& \left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\
=& m^2\sum_{i=1}^{k}\frac{e^{\alpha i}}{\left(1+e^{\alpha i}(p-k)/(epk^4)\right)^2}
\end{aligned}$$

For $n\diamondsuit^2$ we need bounds from both sides. In what follows we write them separately.

$$\begin{aligned}
n\diamondsuit^2(\lambda) \leq& m^2\sum_{i=1}^{k}\frac{e^{\alpha i}}{1}\\
\leq& \frac{m^2k^5}{2\ln(k)},\\
n\diamondsuit^2(\lambda) \geq& \frac{m^2}{\left(1+e^{\alpha k}/(ek^4)\right)^2}\sum_{i=1}^{k}e^{\alpha i}\\
=& \frac{m^2}{(1+e^{-1})^2}\sum_{i=1}^{k}e^{\alpha i}\\
\geq& \frac{m^2k^5}{20\ln(k)}.
\end{aligned}$$

Finally, to satisfy part 4 of Lemma 55 we need

$$n\Diamond^2(\lambda) \geq V(\lambda), \quad n\Lambda^{-1}(\lambda)M(\lambda) \geq a_1\Diamond(\lambda),$$

that is, it is enough to have

$$\frac{m^2k^5}{20\ln(k)} \geq \frac{k}{n} + \frac{n}{p-k}, \quad \frac{m^2k^5}{10\ln(k)} \geq a_1\sqrt{\frac{m^2k^5}{2n\ln(k)}},$$

which is equivalent to

$$\frac{m^2k^5}{20\ln(k)} \geq \left(\frac{k}{n} + \frac{n}{p-k}\right) \vee \frac{50a_1}{n}.$$

For example, we can put

$$m = \frac{b\ln(k)}{k^5}\left(\frac{k}{n} + \frac{n}{p}\right)$$

given that b is a large enough constant. □

B.10 Comparisons with earlier results

Proposition 64. *Assume that $\lambda_i \leq 1$ for any i and $\sum_{i=1}^p \lambda_i \geq \kappa p$ for some constant $\kappa \in (0,1]$. Take $k = 0$, $\lambda = 0$ and some $c > 1$. Suppose additionally that $\kappa p/n \geq \|\boldsymbol{\mu}\|^2 \geq (2ct)^2/(\kappa^2 n)$, and $t^2 < n\kappa$.*

Then

$$\frac{N - ct\Diamond}{[1 + N\sigma_\eta]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n}} \geq \frac{1}{10}\frac{\|\boldsymbol{\mu}\|^2\sqrt{n\kappa}}{\sqrt{p}}.$$

Proof. First of all, due to assumptions on λ_i, λ and k we can write

$$\begin{aligned}
\sum_i \lambda_i^2 \leq& \Lambda,\\
\Lambda =& \sum_i \lambda_i \in [\kappa p, p],\\
V =& \Lambda^{-2}n\sum_{i>k}\lambda_i^2 \leq n/\Lambda \leq \frac{n}{\kappa p},\\
\Delta V \leq& \frac{n\lambda_1^2}{\Lambda^2} + \frac{n\lambda_1^2 + \sum_i \lambda_i^2}{\Lambda^2} \leq \frac{2n}{\kappa^2p^2} + \frac{1}{\kappa p} \leq \frac{3}{\kappa^2 p},\\
\Diamond^2 =& n\Lambda^{-2}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 \leq \frac{n\|\boldsymbol{\mu}\|^2}{\kappa^2p^2},\\
M =& \|\boldsymbol{\mu}\|^2.
\end{aligned}$$

Plugging in those bounds together with $\sigma_\eta < 1$ yields

$$\begin{aligned} n\Lambda^{-1}M - ct\Diamond \geq& \frac{n}{p}\|\boldsymbol{\mu}\|^2 - ct\frac{\sqrt{n}\|\boldsymbol{\mu}\|}{\kappa p}, \\ \left[1 + n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n} \leq& \left[1 + \frac{n}{\kappa p}\|\boldsymbol{\mu}\|^2\right]\sqrt{\frac{n}{\kappa p} + t^2\frac{3}{\kappa^2 p}} + \frac{n\|\boldsymbol{\mu}\|}{\kappa p}. \end{aligned}$$

Next, since $n\|\boldsymbol{\mu}\|^2/p \leq \kappa$ for $t^2 \leq n\kappa$ we can write

$$\left[1 + n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n} \leq [1+1]\sqrt{\frac{4n}{\kappa p}} + \frac{n\|\boldsymbol{\mu}\|}{\kappa p} \leq 5\sqrt{\frac{n}{p\kappa}}.$$

Finally, if $\|\boldsymbol{\mu}\| \geq 2ct/(\kappa\sqrt{n})$, then $\frac{n}{p}\|\boldsymbol{\mu}\|^2 - ct\frac{\sqrt{n}\|\boldsymbol{\mu}\|}{\kappa p} \geq \frac{n}{2p}\|\boldsymbol{\mu}\|^2$. Plugging in that bound in yields the result. □

Proposition 59. *Take $k = 0$ and some $c > 1$. Suppose that $n\lambda_1 < \Lambda$ and $\|\boldsymbol{\mu}\|^2 \geq 2c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$. Then for $t < \sqrt{n}$,*

$$\frac{N - ct\Diamond}{\sqrt{V + t^2\Delta V} + \sqrt{n}\Diamond} \geq \frac{1}{4}\frac{n\|\boldsymbol{\mu}\|^2}{n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} + \sqrt{n}\|\boldsymbol{\Sigma}\|_F + n\|\boldsymbol{\Sigma}\|}. \tag{3.28}$$

Proof. Let's write out the definitions of $\Lambda, M, \Diamond, V, \Delta V$ for the case $k = 0$ with $n\lambda_1 < \lambda + \sum_i \lambda_i$:

$$\begin{aligned} \Lambda =& \lambda + \sum_i \lambda_i = \lambda + \operatorname{tr}(\boldsymbol{\Sigma}), \\ V =& \Lambda^{-2} n \sum_{i>k} \lambda_i^2 = \frac{n\|\boldsymbol{\Sigma}\|_F^2}{(\lambda + \operatorname{tr}(\boldsymbol{\Sigma}))^2}, \\ \Delta V =& \frac{n\lambda_1^2}{\Lambda^2} + \frac{n\lambda_1^2 + \sum_i \lambda_i^2}{\Lambda^2} = \frac{2n\|\boldsymbol{\Sigma}\|^2 + \|\boldsymbol{\Sigma}\|_F^2}{(\lambda + \operatorname{tr}(\boldsymbol{\Sigma}))^2}, \\ \Diamond^2 =& n\Lambda^{-2}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 = \frac{n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2}{(\lambda + \operatorname{tr}(\boldsymbol{\Sigma}))^2}, \\ M =& \|\boldsymbol{\mu}\|^2. \end{aligned}$$

Now we can rewrite our bound as

$$\begin{aligned} & \frac{n\Lambda^{-1}M - ct\Diamond}{\sqrt{V + t^2\Delta V} + \sqrt{n}\Diamond} \\ =& \frac{nM - ct\Lambda\Diamond}{\sqrt{\Lambda^2 V + t^2\Lambda^2\Delta V} + \sqrt{n}\Lambda\Diamond} \\ =& \frac{n\|\boldsymbol{\mu}\|^2 - ct\sqrt{n}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}{\sqrt{n\|\boldsymbol{\Sigma}\|_F^2 + t^2\left(2n\|\boldsymbol{\Sigma}\|^2 + \|\boldsymbol{\Sigma}\|_F^2\right)} + n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}. \end{aligned}$$

We see that for $t \leq \sqrt{n}$ the condition $\|\boldsymbol{\mu}\|^2 \geq 2c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$ ensures that the numerator greater or equal to $n\|\boldsymbol{\mu}_{k:\infty}\|^2/2$. At the same time, the denominator doesn't exceed $n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} + \sqrt{2n}\|\boldsymbol{\Sigma}\|_F + 2n\|\boldsymbol{\Sigma}\|$ Thus, we obtain the desired result. □

Proposition 62. *Take $k = 1$ and some $c > 1$. Assume that $\lambda > 0$, $n\lambda_{k+1} \leq \sum_{i>k} \lambda_i$, $\|\boldsymbol{\mu}_{0:k}\| = 0$, and $\|\boldsymbol{\mu}\|^2 \geq 2c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$. Take any $j > 1$ and define A, B as in Theorem 61. Then for $t \leq \sqrt{n}$*

$$\frac{N - ct\Diamond}{\sqrt{V + t^2\Delta V} + \sqrt{n}\Diamond} \geq \frac{1}{6}\frac{\|\boldsymbol{\mu}\|^2}{A + B + \lambda_j + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}.$$

Proof. Note that under assumption $\|\boldsymbol{\mu}_{0:k}\| = 0$ we have

$$M = \|\boldsymbol{\mu}_{k:\infty}\|^2 = \|\boldsymbol{\mu}\|^2, \quad \Diamond^2 = n\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} = n\Lambda^{-2}\|\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}},$$

and thus, the condition $\|\boldsymbol{\mu}\|^2 \geq 2c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$ can be rewritten as $M \geq 2c\sqrt{n}\Lambda\Diamond$. Therefore, it implies that $n\Lambda^{-1}M - ct\Diamond \geq n\Lambda^{-1}M/2$ for $t \leq \sqrt{n}$. Thus,

$$\frac{n\Lambda^{-1}M - ct\Diamond}{\sqrt{V + t^2\Delta V} + \sqrt{n}\Diamond} \geq \frac{1}{2}\frac{\|\boldsymbol{\mu}\|^2}{n^{-1}\Lambda\sqrt{V + t^2\Delta V} + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}. \tag{B.16}$$

We see that in order to compare Equation (B.16) and (3.32), we need to compare $A + B + \lambda_j$ to $n^{-1}\Lambda\sqrt{V + t^2\Delta V}$ Note that

$$\begin{aligned}
A \geq& \frac{\lambda_1\Lambda}{n\lambda_1 + \Lambda} = n^{-1}\left((n\lambda_1)^{-1} + \Lambda^{-1}\right)^{-1} \geq \frac{1}{2n}\left(\Lambda \wedge n\lambda_1\right),\\
B \geq& \sqrt{\sum_{i\neq 1,j} \lambda_i^2},\\
V =& n^{-1}(1 + n^{-1}\Lambda\lambda_1^{-1})^{-2} + \Lambda^{-2}n\sum_{i>1}\lambda_i^2\\
=& n\Lambda^{-2}\left(n^{-2}(\Lambda^{-1} + n^{-1}\lambda_1^{-1})^{-2} + \sum_{i>1}\lambda_i^2\right)\\
\leq& n\Lambda^{-2}(A^2 + B^2 + \lambda_j^2),\\
\Delta V =& \frac{1}{n} \wedge \frac{n\lambda_1^2}{\Lambda^2} + \frac{n\lambda_2^2 + \sum_{i>1}\lambda_i^2}{\Lambda^2}\\
=& \frac{1}{n\Lambda^2}\left(\Lambda \wedge n\lambda_1\right)^2 + \frac{n^{-1}(n\lambda_2)^2 + \sum_{i>1}\lambda_i^2}{\Lambda^2}\\
\leq& \frac{1}{n\Lambda^2}\left(\Lambda \wedge n\lambda_1\right)^2 + \frac{n^{-1}(n\lambda_1 \wedge \Lambda)^2 + \sum_{i>1}\lambda_i^2}{\Lambda^2}\\
\leq& \frac{1}{n\Lambda^2}\left(2nA\right)^2 + \frac{n^{-1}(2nA)^2 + B^2 + \lambda_j^2}{\Lambda^2}\\
=& n\Lambda^{-2}(8A^2 + B^2/n + \lambda_j^2/n),
\end{aligned}$$

where we used that $n\lambda_2 \leq n\lambda_1$ and $n\lambda_2 \leq \sum_{i>1} \lambda_i \leq \Lambda$ for $\lambda > 0$ to write $n\lambda_2 \leq n\lambda_1 \wedge \Lambda$ when we bounded ΔV. Overall, we get

$$n^{-2}\Lambda^2(V + t^2\Delta V) \leq \frac{1}{n}(A^2 + B^2 + \lambda_j^2) + \frac{t^2}{n}(8A^2 + B^2/n + \lambda_j^2/n),$$

that is, for $t < \sqrt{n}$

$$\begin{aligned} n^{-2}\Lambda^2(V + t^2\Delta V) \leq& 9A^2 + 2B^2/n + 2\lambda_j^2/n, \\ n^{-1}\Lambda\sqrt{V + t^2\Delta V} \leq& 3(A + B + \lambda_j), \end{aligned}$$

which yields the result. □

Proposition 65. *Take real q, r, s such that $0 \leq r < 1 < s$, $0 \leq q < s - r$. Consider $p = n^s$, $\mathbf{\Sigma} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, and $\boldsymbol{\mu} = \sqrt{2\lambda_1/\pi}\boldsymbol{e}_1$, where $\{\lambda_i\}_{i=1}^p$ are given by Equation (3.34). Take $\lambda = 0$, $k = n^r$, and c to be any constant that doesn't depend on n.*

Then, as n goes to infinity, for $t < n^{0.499r}$ the following holds:

$$\frac{N - ct\Diamond}{\sqrt{V + t^2\Delta V} + \sqrt{n}\Diamond} = (1 + o_n(1))\frac{N}{\sqrt{V} + \sqrt{n}\Diamond} = \begin{cases} o_n(1), & 2q + 2r - 1 - s > 0, \\ \frac{1+o_n(1)}{\sqrt{2\pi}} & 2q + 2r - 1 - s = 0, \\ \sqrt{\frac{2}{\pi}} + o_n(1) & 2q + 2r - 1 - s < 0. \end{cases}$$

Here we use $o_n(1)$ to denote quantities that converge to zero as n goes to infinity.

Proof. Throughout the proof we treat q, r, s as constants and use small-oh notation $o(1)$ to denote a function of n, q, r, s that converges to zero as n goes to infinity. Each time we use this notation it denotes a different function.

First of all, let's write out the quantities of interest and plug in the expressions for λ_i.

$$\begin{aligned}
\boldsymbol{\mu} :=&\sqrt{\frac{2\lambda_1}{\pi}}\boldsymbol{e}_1 = \sqrt{2/\pi}n^{(s-q-r)/2}\boldsymbol{e}_1,\\
\Lambda =&\sum_{i>k}\lambda_i = (n^s - n^r)\lambda_{k+1}\\
=&(n^s-n^r)\cdot(1-n^{-q})/(1-n^{r-s})\\
=&n^s - n^{s-q} = n^s(1-o(1)),\\
\Lambda n^{-1}\lambda_1^{-1} =&n^s(1-o(1))\cdot n^{-1}\cdot n^{-s+q+r} = n^{q+r-1}(1-o(1)),\\
\Diamond^2 =&n^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\
=&n^{-1}\cdot(1+\Lambda n^{-1}\lambda_1^{-1})^{-2}\lambda_1^{-1}\cdot\frac{2\lambda_1}{\pi}\\
=&\frac{2+o(1)}{\pi n\left(1+n^{q+r-1}\right)^2},\\
n\Lambda^{-1}M =&\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2\\
=&(1+\Lambda n^{-1}\lambda_1^{-1})^{-1}\lambda_1^{-1}\cdot\frac{2\lambda_1}{\pi}\\
=&\frac{2+o(1)}{\pi(1+n^{q+r-1})},\\
V =&n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right)+\Lambda^{-2}n\sum_{i>k}\lambda_i^2\\
=&n^{-1}n^r(\Lambda n^{-1}\lambda_1^{-1}+1)^{-2}+\Lambda^{-2}n(n^s-n^r)\lambda_{k+1}^2\\
=&\frac{1+o(1)}{n^{1-r}(1+n^{q+r-1})^2}+n^{1-s}(1+o(1)),\\
\Delta V =&\frac{1}{n}\wedge\frac{n\lambda_1^2}{\Lambda^2}+\frac{n\lambda_{k+1}^2+\sum_{i>k}\lambda_i^2}{\Lambda^2}\\
=&n^{-1}\wedge n^{1-2r-2q}(1+o(1))+n^{-s}(1+o(1)).
\end{aligned}$$

Now let's plug this into the quantity of interest. Note that as long as $t=o(\sqrt{n})$, we have $t\Diamond = o(n\Lambda^{-1}M)$. Moreover, $\Delta V/V = O(n^{-r})$, indeed

$$n^r\Delta V = \left(\frac{1}{n^{1-r}\vee n^{2(q+r-1)-r}}+n^{r-s}\right)(1+o(1))\le V(1+o(1)),$$

since $r<1$. Thus, if $t^2=o(n^r)$, then $V+t^2\Delta V = V(1+o(1))$. Now note that $n\Lambda^{-1}M = \Diamond\sqrt{n}(\sqrt{2/\pi}+o(1))$. That is,

$$\frac{n\Lambda^{-1}M}{\sqrt{V}+\sqrt{n}\Diamond} = \frac{\sqrt{2/\pi}+o(1)}{1+\sqrt{V/(n\Diamond^2)}}.$$

So, the only thing left is to compare V and $n\Diamond^2$:

$$\frac{V}{n\Diamond^2} = \frac{\pi + o(1)}{2}\left(n^{r-1} + n^{1-s}(1 + n^{q+r-1})^2\right).$$

Since $r < 1 < s$, this ratio goes to infinity if and only if $n^{1-s}(n^{q+r-1})^2$ goes to infinity, that is $2q + 2r - 1 - s > 0$, which yields the result. □