Journal of Machine Learning Research 27 (2026) 1-121 Submitted 3/25; Revised 9/26; Published 9/26

Benign Overfitting and the Geometry of the Ridge Regression Solution in Binary Classification

Alexander Tsigler ALEXANDER_TSIGLER@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
367 Evans Hall, Berkeley, CA 94720-3860

Luiz F. O. Chamon LUIZ.CHAMON@POLYTECHNIQUE.EDU
Department of Applied Mathematics
École polytechnique, Institut Polytechnique de Paris
91128 Palaiseau CEDEX

Spencer Frei SFREI@GOOGLE.COM
Google DeepMind
1600 Amphitheatre Parkway, Mountain View, CA 94043

Peter L. Bartlett PETER@BERKELEY.EDU
Departments of Statistics and EECS
University of California, Berkeley and Google DeepMind
367 Evans Hall, Berkeley, CA 94720-3860

Editor: Daniel Roy

Abstract

In this work, we investigate the behavior of ridge regression in an overparameterized binary classification task. We assume examples are drawn from (anisotropic) class-conditional cluster distributions with opposing means and we allow for the training labels to have a constant level of label-flipping noise. We characterize the classification error achieved by ridge regression under the assumption that the covariance matrix of the cluster distribution has a high effective rank in the tail. We show that ridge regression has qualitatively different behavior depending on the scale of the cluster mean vector and its interaction with the covariance matrix of the cluster distributions. In regimes where the scale is very large, the conditions that allow for benign overfitting turn out to be the same as those for the regression task. We additionally provide insights into how the introduction of label noise affects the behavior of the minimum norm interpolator (MNI). The optimal classifier in this setting is a linear transformation of the cluster mean vector and in the noiseless setting the MNI approximately learns this transformation. On the other hand, the introduction of label noise can significantly change the geometry of the solution while preserving the same qualitative behavior.

Keywords: benign overfitting, overparameterized models, ridge regression, linear classification, binary classification, minimum norm interpolator

©2026 Alexander Tsigler, Luiz F. O. Chamon, Spencer Frei, and Peter L. Bartlett.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-0531.html.

Contents

1 Introduction **4**
1.1 Binary classification problem setting 4
1.2 Assumptions on the distribution of the covariates 5
1.3 First result: recovering the geometry in the noiseless setting 6
1.4 Second result: benign overfitting 7
1.5 Comparison to the regression problem 8
1.6 Related work 9
1.7 Structure of the paper 11
1.8 Additional notation 11

2 Geometric picture for minimum norm interpolation **12**
2.1 MNI without label-flipping noise 13
2.2 MNI with label-flipping noise and benign overfitting 14

3 Assumptions on the data **15**
3.1 Gram matrix of the tails 15
3.2 Algebraic assumptions 16

4 Main results **17**
4.1 Lower bound 20
4.2 Upper bound 20
4.3 Tight bound for a quantile 21
4.4 Regimes of the lower bound 22
4.4.1 MNI without label-flipping noise 22
4.4.2 MNI with label-flipping noise 24
4.5 Benign overfitting 28

5 Effect of regularization **29**
5.1 Increasing regularization never helps in the noiseless case 30
5.2 Increasing regularization does nothing in some regimes 32
5.3 Increasing regularization may cause harm by breaking the balance between the tail and the spiked part 32
5.4 Increasing regularization can harm by destroying "recovery of the geometry" 34
5.5 Regularization with label-flipping noise 36

6 Detailed comparison with previous results **37**
6.1 Comparison with (Cao et al., 2021) 37
6.2 Comparison with (Wang and Thrampoulidis, 2021) 38
6.3 Comparison with (Chatterji and Long, 2021) 40
6.4 Comparison with (Muthukumar et al., 2021) 41

7 Conclusions and further directions **42**

A Notation **44**
A.1 Abbreviations . . . 44
A.2 Basic notation . . . 44
A.3 Problem setting . . . 45
A.4 Ridge regression and minimum norm interpolating solutions . . . 45
A.5 Objects from geometric derivations . . . 46
A.6 Separation of the first k components . . . 46
A.7 High probability events . . . 46
A.8 Quantities from the main bounds . . . 47
A.9 Alternative forms of main quantities . . . 47

B Proof outline and different forms of the main quantities **47**
B.1 Lower bound proof sketch . . . 47
B.2 Upper bound proof sketch . . . 48

C Formulas for the solution **50**

D General probabilistic results **55**

E Some important relations **59**

F Randomness in labels **66**

G Algebraic decompositions **69**
G.1 Techniques and proof strategy . . . 70
G.2 $\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|$ . . . 71
G.3 $\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp}$ . . . 72
G.3.1 Bound from below . . . 72
G.3.2 Bound from above . . . 74
G.4 $\|\boldsymbol{\mu}_{\perp}\|_{\boldsymbol{\Sigma}}^2$ . . . 75
G.5 $\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})$ . . . 76
G.6 $\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\|$ . . . 77

H Randomness in covariates **78**

I Proof of the main lower bound **81**

J Proof of tightness **86**
J.1 Numerator . . . 87
J.2 Denominator . . . 90
J.3 The ratio . . . 96

K Analysis of ridge regularization **99**

L Comparisons with earlier results **115**

1. Introduction

Following empirical observations that large neural networks can achieve vanishing error on noisy training data and still perform well on test data, there has been a plethora of work aimed at theoretically understanding the generalization performance of overparameterized models (see Bartlett et al. (2021) and Belkin (2021) for relevant surveys). In particular, there is significant interest in understanding when interpolating noisy data (i.e., *overfitting*) can be *benign*, i.e., when an estimator can achieve near-zero training error while simultaneously generalizing well. The model class, the choice of the estimation procedure, and properties of the training data as well as the population distribution are key factors that interact to influence the consequences of overfitting.

To date, the most complete picture we have of this *benign overfitting* phenomenon is for the minimum ℓ_2-norm solution of well-specified linear regression. Here, this phenomenon arises due to a self-induced regularization coming from a population covariance matrix with large effective rank (relative to the number of samples) (Bartlett et al., 2020). Beyond linear regression, however, our understanding is far less developed, even in the setting of linear classification. The main goal of this paper is to characterize how benign overfitting can occur in this setting.

1.1 Binary classification problem setting

We consider a mixture of two classes with the same covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ 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 by

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

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 $\boldsymbol{Q} := \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$—the matrix of covariates with centers of clusters subtracted. We consider the overparameterized regime, that is $p > n$, and linear classifiers that assign the label $\operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x})$ to the point $\boldsymbol{x} \in \mathbb{R}^p$. Here, $\boldsymbol{w} \in \mathbb{R}^p$ is the weight vector of the classifier.

Denote by $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ the function that provides control over the deviations of the rows of $\boldsymbol{Z}$ in all directions, namely,

$$\phi(t) := \sup_{\boldsymbol{v}\in\mathcal{S}^{p-1}} \mathbb{P}(\boldsymbol{z}^\top \boldsymbol{v} < -t),$$

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

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

where we introduced $\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}} := \sqrt{\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}}$. That is, the quantity $\boldsymbol{\mu}^\top\boldsymbol{w}/\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}$ controls the probability of predicting the wrong label on a new data point. Moreover, if rows of $\boldsymbol{Z}$ have a 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 paper are bounds on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, where $\boldsymbol{w}_{\text{ridge}}$ is the solution of the (noisy) ridge regression problem defined as follows. First, we assume we are given a vector of labels $\hat{\boldsymbol{y}}$ perturbed by 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

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

where $\boldsymbol{I}_n \in \mathbb{R}^{n\times n}$ is the 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, the solution in (1) exactly interpolates the labels $\hat{\boldsymbol{y}}$. When $\lambda = 0$, we introduce a separate notation for $\boldsymbol{w}_{\text{ridge}}$, namely, $\boldsymbol{w}_{\text{MNI}}$. Here, MNI stands for the *Minimum Norm Interpolating* solution.

Throughout the paper, we will also use the notation $\boldsymbol{A} := \lambda \boldsymbol{I}_n + \boldsymbol{Q}\boldsymbol{Q}^\top$ and $\boldsymbol{\nu} := \boldsymbol{Q}\boldsymbol{\mu}$.

1.2 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 Tsigler and Bartlett (2023) that generalizes (Bartlett et al., 2020).

First, 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, $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$. The remainder of the paper considers this basis. We assume that for some k—that 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 3.1. When the data is Gaussian (or, more generally, if the matrix $\boldsymbol{Z}$ has independent sub-Gaussian elements), that condition would be

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

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 level of regularization applied. In (Bartlett et al., 2020) it was shown that for $\lambda = 0$, condition (2) is necessary for benign overfitting to happen in linear regression.

In view of this separation of the first k components, we introduce the following notation taken from Tsigler and Bartlett (2023): for any $k \in \{0, 1, \dots, p\}$ and any matrix $\boldsymbol{M} \in \mathbb{R}^{n\times p}$, we denote by $\boldsymbol{M}_{0:k}$ the matrix comprised of the first k columns of $\boldsymbol{M}$.[1] Analogously, we denote by $\boldsymbol{M}_{k:\infty}$ the matrix comprised of the last $p-k$ columns of $\boldsymbol{M}$. For any $\boldsymbol{u} \in \mathbb{R}^p$, we denote by $\boldsymbol{u}_{0:k}$ the vector comprised of the first k components of $\boldsymbol{u}$ and by $\boldsymbol{u}_{k:\infty}$ 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" and to the remaining components as the "tail of the covariance."

Apart from the effective rank condition described above, we also need several concentration inequalities to hold. Those inequalities, however, are rather standard (such as law

1. When $k = 0$, this matrix is empty and all the terms involving the indices $0 : k$ become zero.

of large numbers for i.i.d. random variables or sample covariance concentration in low dimensions), so we choose to assume that those inequalities hold directly instead of deriving them from assumptions on the distribution of the data. We introduce these inequalities in Section 3.2.

Our motivation to consider such a regime was initially technical: we believed that the quantities of interest could be accurately evaluated in this regime. However, our results suggest that such a structure of the data is necessary for benign overfitting to occur in classification. We elaborate on this point in Section 7.

1.3 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 us consider those two terms separately. On the one hand, the vector $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ has a symmetric distribution and no dependence on $\boldsymbol{\mu}$, the class centers, thus playing the role of a noise term. The vector $\boldsymbol{\mu}$ should therefore be large enough, i.e., the classes should be sufficiently separated, for this 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 vector $\boldsymbol{w}$ that minimizes $\boldsymbol{\mu}^\top \boldsymbol{w}/\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}$ is proportional to $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Here, however, $\boldsymbol{w}_{\text{ridge}}$ incorporates the tail of the covariance $\sum_{i>k}\lambda_i$ in addition to the explicit ridge regularization λ. Another way to think about this is to introduce, as in (Bartlett et al., 2020),

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

We then obtain, up to a constant factor,

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

Notice that the ridge regression solution performs the optimal linear transformation in the first k^* coordinates, but not in the remaining ones. Throughout this paper, we refer to this effect as "recovering the geometry" in the first k^* components.

Based on these observations, we show that there are three regimes of interest for the noiseless setting. When the magnitude of $\boldsymbol{\mu}$ is small, $\boldsymbol{w}_{\text{ridge}}$ behaves as the "noise term" $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ in both $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ and $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$. The quantity $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ will be negative with probability close to 50%, resulting in no meaningful bound on the classification performance. As the magnitude of $\boldsymbol{\mu}$ grows, the term $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ starts dominating in $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$, but not in $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$. This regime already yields a non-vacuous classification guarantee, but still does not exhibit the full "recovery of the geometry" behavior. Finally, when $\boldsymbol{\mu}$ becomes large enough, $\boldsymbol{w}_{\text{ridge}}$ performs effectively as $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$.

For the special case of zero regularization (that is, $\lambda = 0$ and $\boldsymbol{w}_{\text{ridge}} = \boldsymbol{w}_{\text{MNI}}$), Section 2.1 provides geometric derivations for the decomposition of the solution vector into the "noise" and "signal" components as stated above. We further explore the quantitative implications

of this decomposition in Section 4.4.1. It is worth noting that $(\boldsymbol{\Sigma}+n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ 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 relative magnitude of the noise term $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$. Nevertheless, in Section 5 we show that it is indeed the case and that one cannot gain a significant (in a certain sense) improvement in performance by increasing λ beyond the point at which the data has high effective rank in the tail of the covariance (in the absence of label-flipping noise, i.e., $\eta = 0$). Moreover, we show that depending on the relation between $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$, decreasing the explicit regularization λ even below zero can significantly improve performance. We demonstrate two scenarios in which such improvements occur: in the first one, the balance between the solution in the first k^* components and that in the remaining ones becomes advantageous (Section 5.3). In the second one, decreasing regularization increases k^* and the performance improves by "recovering the geometry" in more coordinates (Section 5.4).

1.4 Second result: benign overfitting

In the presence of label-flipping noise ($\eta > 0$), we show that the structure of the solution vector can be significantly different from the noiseless case depending on the magnitude of $\boldsymbol{\mu}$. For MNI (Section 2.2), adding label-flipping noise scales the noiseless solution by a random factor and aggregates an additional "noise component" (that has no dependence on $\boldsymbol{\mu}$) in the orthogonal direction. As $\boldsymbol{\mu}$ becomes large, the random scaling factor 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" begins to dominate over the noiseless solution in norm.

Yet, 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 similar. Specifically, the bound on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}$ remains practically the same and the bound on $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ picks up an additional term corresponding to the magnitude of the new orthogonal "noise component." As a result, our bounds suggest[2] that the noisy solution goes through the same regimes as the noiseless one, with an additional regime when $\boldsymbol{\mu}$ is very large in magnitude. In the latter, our bound loses the dependence on $\boldsymbol{\mu}$ and becomes a function only of the covariance. Interestingly, the conditions under which the ratio $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ becomes large, that is, the conditions under which benign overfitting occurs, are exactly the same as those from the regression literature (Bartlett et al., 2020; Tsigler and Bartlett, 2023).

A comprehensive discussion of the regimes through which the bound goes in the noisy case can be found in Section 4.4.2. In Section 4.5 we present a quantitative result that gives sufficient conditions for benign overfitting in the binary classification setting we consider.

2. Unfortunately, we do not provide a matching upper bound on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ for the case with label-flipping noise, so we can only make claims about the bound rather than the ratio itself. Nevertheless, we do believe this statement also applies to $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ and provide a rationale as to why in Section 7.

1.5 Comparison to the regression problem

As we mentioned earlier, our setting and results parallel those from the papers on benign overfitting in linear and ridge regression (Bartlett et al., 2020; Tsigler and Bartlett, 2023). In this section, we recap their main conclusions and highlight some connections and differences with our classification setup.

Bartlett et al. (2020) and Tsigler and Bartlett (2023) consider the minimum-norm interpolating solution for an overparameterized linear regression problem. In our notation, it can be formulated as

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

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 $\hat{\boldsymbol{\theta}}$ has a closed form expression, namely,

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

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

$$\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\right\|_{\boldsymbol{\Sigma}} \leq \left\|(\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\theta}^*\right\|_{\boldsymbol{\Sigma}} + \left\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{\varepsilon}\right\|_{\boldsymbol{\Sigma}}. \tag{3}$$

The first term on the right-hand side of (3) constitutes the bias of the MNI regression solution and the second term constitutes its variance. The main result of Bartlett et al. (2020) is that under the structure that we introduced in Section 1.2, the variance term becomes small. This occurs because in the first k components, the noise vector $\boldsymbol{\varepsilon}$ gets projected from dimension n onto a small dimension k, so that its energy gets damped by a factor k/n. The remainder of the noise vector is then 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 the components $k:\infty$, so that the fact that they absorbed the noise is irrelevant. When it comes to the bias term, Tsigler and Bartlett (2023) show that no learning happens in the components $k:\infty$, that is, almost all the energy of the signal $\|\boldsymbol{\theta}^*_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}$ goes into the bias term. In the first k components, on the other hand, the bias term behaves as in classical k-dimensional ridge regression 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 is not relevant for predictions.

In contrast, the classification setting considered in this paper is fundamentally different. Indeed, in regression, the "signal vector" $\boldsymbol{\theta}^*$ is an element of a dual space measured through the data matrix $\boldsymbol{Q}$. 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 $\boldsymbol{Q}$ 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 regression papers 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 $\|\boldsymbol{\mu}_{0:k}\| = 0$), this would result in high classification error since no learning happens in the tail. This is not correct, as our bounds imply arbitrarily high classification accuracy in this setting. Since $\boldsymbol{\mu}$ is baked into $\boldsymbol{X}$, a plausible hypothesis could be 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 our derivations.

Yet, our argument shows very strong connections to the regression setting. In Section 1.3, we stated that in the noiseless case, $\boldsymbol{w}_{\text{ridge}}$ behaves as $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$. Observe that the first term in that expression is almost exactly the variance part of the regression solution [last term in (3)]. Moreover, the vector $(\boldsymbol{\Sigma} + n^{-1}\Lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ arises as an approximation of 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 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 our paper 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. In the presence of label-flipping noise (Section 1.4), we mention an additional "orthogonal noise component." As it turns out, this component has a very similar structure to $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ (see Section 2.2 for the precise derivation). Hence, in the large $\boldsymbol{\mu}$ regime with label-flipping noise, our bound on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}} / \|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ has the form $1/\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}$. 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 suggests that the mechanism by which the regression solution "hides" the noise it interpolates is similar to that used by the classification solution.

Overall, though we identify very concrete connections between our classification results and the regression results from Bartlett et al. (2020) and Tsigler and Bartlett (2023), identifying either a clear unifying picture of these two settings or a fundamental difference between them remains an intriguing open question.

1.6 Related work

Despite the fact that the literature on linear classification in high dimensions is vast, only a few papers studied problems with general covariance structure and the impact of that structure on the prediction accuracy. Here, we only review those works, referring the reader to Loureiro et al. (2021) and Wang et al. (2021) for a broader review of the related literature.

Existing results can be split into asymptotic and non-asymptotic. 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.

Montanari et al. (2023) 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. Loureiro et al. (2021) considers 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 the 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 that can accurately predict properties of certain large stochastic systems. Huang and Yang (2021) 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.

Shamir (2022) considers a data structure 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.

These papers, however, are substantially different from this work. They either consider the maximum margin solution instead of the ridge regression solution or obtain results 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 theirs. The works we discuss in the remainder of this section, however, turn out to be directly comparable to ours.

Still in the asymptotic literature, Muthukumar et al. (2021) considers a model similar to that of Montanari et al. (2023), but with a different choice of the covariance structure. Their main result is given for a certain "bi-level" covariance whose eigenvalues 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 1.2. Even though the goal of Muthukumar et al. (2021) is to study the maximum margin solution, the approach they take is motivated by a recent observation that under certain assumptions, the maximum margin solution coincides with the MNI solution (Hsu et al., 2021; Ardeshir et al., 2021). This phenomenon is known as "support proliferation." Because of that, the main result of Muthukumar et al. (2021) is actually derived for the minimum norm interpolating solution, which makes it possible to compare it to our result.

When it comes to the non-asymptotic literature, Cao et al. (2021) study the classification error of the MNI solution and Wang and Thrampoulidis (2021) study the ridge regression solution, both motivated by the idea of support proliferation. Finally, Chatterji and Long (2021) obtain a bound on the misclassification probability of the maximum margin solution in binary classification. To approach the maximum margin solution, they use its characterization for separable data as the limit of gradient descent on logistic loss. As it turns out, however, the assumptions they consider imply that support proliferation happens on the event when their proof works.

We give detailed comparisons with Cao et al. (2021); Wang and Thrampoulidis (2021); Chatterji and Long (2021); Muthukumar et al. (2021) in Section 6. Interestingly, support

proliferation requires the whole data distribution to have high effective rank, so most of the results from those papers correspond to our result with $k = 0$.

Finally, we note that a version of our work was previously included as part of the first author's PhD thesis (Tsigler, 2024).

1.7 Structure of the paper

We start in Section 2 by considering MNI and providing geometric descriptions of the solution vector. In Section 3, we introduce our assumptions on the distribution of the data to obtain quantitative bounds. Section 4 gives those quantitative results and explains the regimes they go through depending on the magnitude of $\boldsymbol{\mu}$. We put all the technical steps of the proofs of those bounds in the appendix. In Section 5, we study the influence of ridge regularization on the error of the classifier. Finally, we provide detailed comparisons with the previous literature in Section 6.

1.8 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 $a+1$. 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 $\boldsymbol{e}_i$. 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 at least a large (at most a small) constant times b." More precisely, when a part of an assumption (e.g., in a statement of the form "if $a \gg b$ then ..."), $a \gg b$ means that there is a large enough constant C such that if $a > Cb$ then the conclusion holds. When a part of a conclusion (e.g., "if ... then $a \gg b$"), $a \gg b$ means that $a > Cb$, and the constant C can be arbitrarily large if the constants in the assumption are taken large enough. For example, the statement "if $a \gg b$ then $c \gg d$" means that for any constant $C_{c,d} > 0$ there exists a constant $C_{a,b} > 0$ such that if $a > C_{a,b}b$, then $c > C_{c,d}d$.

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 elements of $\boldsymbol{v}$. For any positive integer d, we denote by $\mathbf{0}_d \in \mathbb{R}^d$ the vector of all zeros, by $\mathbf{1}_d \in \mathbb{R}^d$ the vector of all ones, and by $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ the identity matrix. 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 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 the symmetric matrix $\boldsymbol{M}$. 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 define $\|\boldsymbol{u}\|_{\boldsymbol{M}} := \sqrt{\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u}}$. For any matrix $\boldsymbol{M} \in \mathbb{R}^{n\times p}$, we denote its pseudo-inverse by $\boldsymbol{M}^\dagger \in \mathbb{R}^{p\times n}$.

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

For readers' convenience we put a table of notation in Appendix A.

2. Geometric picture for minimum norm interpolation

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

Explicitly, the MNI solution can be defined as

$$\boldsymbol{w}_{\text{MNI}} := \arg\min_{\boldsymbol{w}\in\mathbb{R}^p} \|\boldsymbol{w}\| \text{ s.t. } \boldsymbol{X}\boldsymbol{w} = \hat{\boldsymbol{y}}, \tag{4}$$

i.e., the vector with minimum Euclidean norm that interpolates the (potentially noisy) labels. There is an explicit formula for (4), namely $\boldsymbol{w}_{\text{MNI}} = \boldsymbol{X}^\dagger \hat{\boldsymbol{y}}$, where $\boldsymbol{X}^\dagger$ denotes the pseudo-inverse of $\boldsymbol{X}$. Unfortunately, this characterization is not convenient for us since we want to decouple the contributions of $\boldsymbol{\mu}$ from $\boldsymbol{Q}$. An alternative, more amenable definition can be obtained from the dual form of the convex optimization problem (4), namely,

$$\tilde{\boldsymbol{w}}_{\text{MNI}} := \boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}} \boldsymbol{\alpha}, \text{ where } \boldsymbol{\alpha} = \text{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^n} \|\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}} \boldsymbol{\alpha}\| \text{ s.t. } \boldsymbol{\alpha}^\top \mathbf{1}_n = 1, \tag{5}$$

where $\boldsymbol{D}_{\hat{\boldsymbol{y}}} := \text{diag}(\hat{\boldsymbol{y}})$. 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 the affine span is the set of linear combinations whose coefficients sum to one, i.e., the set $\{\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}} \boldsymbol{\alpha} : \boldsymbol{\alpha}^\top \mathbf{1}_n = 1\}$. The precise result is given by the following proposition.

Proposition 1 *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:*

$$\boldsymbol{w}_{MNI} = \frac{\tilde{\boldsymbol{w}}_{MNI}}{\|\tilde{\boldsymbol{w}}_{MNI}\|^2}, \quad \tilde{\boldsymbol{w}}_{MNI} = \frac{\boldsymbol{w}_{MNI}}{\|\boldsymbol{w}_{MNI}\|^2}. \tag{6}$$

Proof We provide here a geometric argument. Alternatively, the proof can be carried out using Lagrangian duality to relate (4) and (5). First, because $\hat{\boldsymbol{y}} \in \{-1, 1\}^n$, we can rewrite the definition of MNI as

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

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. As for the norm, notice that since $\tilde{\boldsymbol{w}}_{\text{MNI}}$ belongs to the affine span, it must be that $\boldsymbol{w}_{\text{MNI}}{}^\top \tilde{\boldsymbol{w}}_{\text{MNI}} = 1$. This yields (6). ■

Since both $\boldsymbol{w}_{\text{MNI}}$ and $\tilde{\boldsymbol{w}}_{\text{MNI}}$ have the same direction, they result in the same classification rule. Nevertheless, it is sometimes more convenient to work with one or the other. Indeed, $\tilde{\boldsymbol{w}}_{\text{MNI}}$ is more suitable to study the noiseless case (Section 2.1) whereas $\boldsymbol{w}_{\text{MNI}}$ turns out to be better suited in the presence of label-flipping noise (Section 2.2).

2.1 MNI without label-flipping noise

In this section, we study the case without label-flipping noise, that is, $\eta = 0$ and $\hat{\boldsymbol{y}} = \boldsymbol{y}$. In this case, we write (4) and (5) as

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

Here, the superscript c stands for "clean." We start by using the expression for $\boldsymbol{X}$ to get

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

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 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 $\boldsymbol{w}^c_{\text{MNI}}$ and $\tilde{\boldsymbol{w}}^c_{\text{MNI}}$.

Let us start with the case $\boldsymbol{\mu} = \mathbf{0}_p$ and then see the effect of adding $\boldsymbol{\mu}$. When $\boldsymbol{\mu} = \mathbf{0}_p$, the matrix $\boldsymbol{X}$ coincides with $\boldsymbol{Q}$, so that

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

Note that $\|\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\mathbf{0}_p)\|^2 = (\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^{-1}$. Hence, as we add $\boldsymbol{\mu}$, the columns of $\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$ are shifted 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 parallel to the affine span of those columns as $\mathcal{Q}_A$. Note that $\mathcal{Q}_A$ is orthogonal to $\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\mathbf{0}_p)$ and that $\mathcal{Q} = \mathcal{Q}_A \oplus \langle \tilde{\boldsymbol{w}}^c_{\text{MNI}}(\mathbf{0}_p)\rangle$—the direct sum of $\mathcal{Q}_A$ and the line spanned by $\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\mathbf{0}_p)$. Thus, we can decompose $\boldsymbol{\mu}$ into 3 orthogonal components: one perpendicular to $\mathcal{Q}$, one lying in $\mathcal{Q}_A$, and one in the direction of $\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\mathbf{0}_p)$. Explicitly,

$$\boldsymbol{\mu} = \boldsymbol{\mu}_\perp + \boldsymbol{\mu}_{\|\mathcal{Q}_A} + \frac{\boldsymbol{\mu}^\top \tilde{\boldsymbol{w}}^c_{\text{MNI}}(\mathbf{0}_p)}{\|\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\mathbf{0}_p)\|^2}\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\mathbf{0}_p). \tag{7}$$

Recall that $\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\boldsymbol{\mu})$ 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 in the direction of $\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\mathbf{0}_p)$ doesn't change the linear span, but it shifts the affine span orthogonally and is therefore added to the projection. Finally, $\boldsymbol{\mu}_\perp$ shifts the linear span orthogonally, so it also gets added to the projection. Therefore, we get

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

Plugging in the expressions for each variable, we get the formula

$$\tilde{\boldsymbol{w}}^c_{\text{MNI}}(\boldsymbol{\mu}) = \frac{\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}} + \underbrace{(\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu}}_{\boldsymbol{\mu}_\perp} + \frac{\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}, \tag{8}$$

where we use the notation $\boldsymbol{\nu} := \boldsymbol{Q}\boldsymbol{\mu}$.

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

In Section 1.3, we claimed that the noiseless 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 $(n\Lambda^{-1}\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$ and $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \approx n\Lambda^{-1}$. Additionally, 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. That domination can be expected from our geometric derivation: the second term, $\boldsymbol{\mu}_\perp$, is a projection of $\boldsymbol{\mu}$ onto $\mathcal{Q}^\perp$, while its sum with the third term is the projection of $\boldsymbol{\mu}$ onto $\mathcal{Q}_A^\perp$—a random subspace which differs by only one dimension. We explain these quantitative implications in Section 4.4.1.

2.2 MNI with label-flipping noise and benign overfitting

Let us proceed with the case where labels are contaminated by label-flipping noise. Note that we now use the MNI characterization $\boldsymbol{w}_{\text{MNI}}$ in (4) rather than $\tilde{\boldsymbol{w}}_{\text{MNI}}$ from (5).

Start by recalling that we denoted the linear span of the columns of $\boldsymbol{Q}^\top \boldsymbol{D}_{\boldsymbol{y}}$ as $\mathcal{Q}$. For any $\boldsymbol{v} \in \mathbb{R}^p$, let the projection of $\boldsymbol{v}$ on $\mathcal{Q}$ be $\boldsymbol{v}_\parallel$ and the projection of $\boldsymbol{v}$ on $\mathcal{Q}^\perp$ be $\boldsymbol{v}_\perp$. The main idea is to separately consider $\boldsymbol{w}_{\text{MNI}\perp}$ and $\boldsymbol{w}_{\text{MNI}\parallel}$. Firstly, since $\boldsymbol{w}_{\text{MNI}}$ lies in the span of the columns of $\boldsymbol{X}^\top = \boldsymbol{Q}^\top + \boldsymbol{\mu}\boldsymbol{y}^\top$, its projections onto $\mathcal{Q}^\perp$ must be collinear with $\boldsymbol{\mu}_\perp$. In other words, $\boldsymbol{w}_{\text{MNI}\perp} = \alpha \boldsymbol{\mu}_\perp$ for some scalar α.

Secondly, note that $\boldsymbol{Q}\boldsymbol{v} = \boldsymbol{0}_n$ for any $\boldsymbol{v} \in \mathcal{Q}^\perp$, so that we have $\boldsymbol{X}\boldsymbol{v} = \boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{v} \propto \boldsymbol{y}$. In particular, this implies that $\boldsymbol{w}_{\text{MNI}\perp}$ always interpolates labels that are proportional to the "clean" labels $\boldsymbol{y}$, while $\boldsymbol{w}_{\text{MNI}\parallel}$ must interpolate the residuals. In other words, for a scalar β and another scalar $\beta' = (\beta + \boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}\parallel})$, it holds that

$$\begin{aligned}\boldsymbol{X}\boldsymbol{w}_{\text{MNI}\parallel} =& \hat{\boldsymbol{y}} - \beta \boldsymbol{y},\\ \boldsymbol{Q}\boldsymbol{w}_{\text{MNI}\parallel} =& \boldsymbol{X}\boldsymbol{w}_{\text{MNI}\parallel} - \boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}\parallel} = \hat{\boldsymbol{y}} - \beta' \boldsymbol{y}.\end{aligned}$$

Observe now that, since $\mathcal{Q}$ is the span of rows of $\boldsymbol{Q}$, any vector $\boldsymbol{v} \in \mathcal{Q}$ is uniquely defined by its scalar products with the rows of $\boldsymbol{Q}$, and that the inverse mapping is linear: $\boldsymbol{v} = \boldsymbol{Q}^\dagger(\boldsymbol{Q}\boldsymbol{v}) = \boldsymbol{Q}^\top \boldsymbol{A}^{-1}(\boldsymbol{Q}\boldsymbol{v})$. Thus $\boldsymbol{w}_{\text{MNI}\parallel} = \boldsymbol{Q}^\top \boldsymbol{A}^{-1}(\hat{\boldsymbol{y}} - \beta' \boldsymbol{y})$. That is, $\boldsymbol{w}_{\text{MNI}\parallel}$ lies in the span of $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}$ and $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$. Since we are dealing with the minimum norm solution, it is convenient to orthogonalize that basis. It is straightforward to check that $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}(\hat{\boldsymbol{y}} - \xi \boldsymbol{y})$ is orthogonal to $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ for $\xi = \frac{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}$. By letting $\tilde{\boldsymbol{y}} := \hat{\boldsymbol{y}} - \xi \boldsymbol{y}$, we then see that $\boldsymbol{w}_{\text{MNI}\parallel}$ lies in the two-dimensional subspace for which $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ and $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ form an orthogonal basis. Combining these arguments, we see that for scalars α and γ we have

$$\boldsymbol{w}_{\text{MNI}} = \boldsymbol{w}_{\text{MNI}\perp} + \boldsymbol{w}_{\text{MNI}\parallel} = \alpha \boldsymbol{\mu}_\perp + \gamma \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}.$$

To determine the values of α and γ, observe that all three vectors on the right-hand side of the expression are orthogonal to each other and that $\boldsymbol{w}_{\text{MNI}}$ is, by definition, the minimum norm vector that interpolates labels $\hat{\boldsymbol{y}}$. Since the term $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ interpolates labels $\tilde{\boldsymbol{y}} + \boldsymbol{y}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$, α and γ must be such that $\alpha \boldsymbol{\mu}_\perp + \gamma \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ has the minimum possible norm in the span of $\boldsymbol{\mu}_\perp$ and $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ while interpolating the labels $\hat{\boldsymbol{y}} - \tilde{\boldsymbol{y}} - \boldsymbol{y}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}} =$

$(\xi - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}})\boldsymbol{y}$. But recall that $\boldsymbol{w}^c_{\text{MNI}}$ has the minimum possible norm among all vectors that interpolate $\boldsymbol{y}$ and also happens to lie in the span of $\boldsymbol{\mu}_\perp$ and $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$. Hence, it must be that $\alpha\boldsymbol{\mu}_\perp + \gamma \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} = (\xi - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}})\boldsymbol{w}^c_{\text{MNI}}$, that is,

$$\boldsymbol{w}_{\text{MNI}} = \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}} + (\xi - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}})\boldsymbol{w}^c_{\text{MNI}}. \tag{9}$$

We provide a comprehensive discussion of the quantitative results regarding $\boldsymbol{w}_{\text{MNI}}$ in the noisy case when η is a small constant in Section 4.4.2. In that section, we will see that in Euclidean norm, the term $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ always dominates the norm of $\boldsymbol{w}_{\text{MNI}}$. Note that $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ acts as a "noise term" as it has no dependence on $\boldsymbol{\mu}$ and has a symmetric distribution. Moreover, though $\xi \approx 1$, it turns out that $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ also has a symmetric distribution, given that it is a linear function of $\tilde{\boldsymbol{y}}$. Since $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ is linear in $\boldsymbol{\mu}$, when $\boldsymbol{\mu}$ is large, the term $(\xi - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}})\boldsymbol{w}^c_{\text{MNI}}$ points in the opposite direction of the noiseless solution with probability close to 0.5. Therefore, the noisy solution looks significantly different than the noiseless to an observer who only looks at $\boldsymbol{w}_{\text{MNI}}$ in Euclidean geometry without considering the structure of $\boldsymbol{\Sigma}$.

Despite those differences, the bound for the noisy case is surprisingly similar to the bound in the noiseless case. Due to certain algebraic cancellations, the formula for $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}$ in the noisy case turns out to be almost the same as in the noiseless case and the bound does not change. Unfortunately, we do not have an intuitive geometric explanation as to why that happens. When it comes to $\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$, according to our bounds, $\|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$ is dominated by $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$. As a result, only one additional term appears in the bound: when $\boldsymbol{\mu}$ is large, $\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \approx \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}$. This leads to the "benign overfitting" regime that we announced in Section 1.4.

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. Those bounds, in their turn, require assumptions. In this section, we explain the 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.

3.1 Gram matrix of the tails

The central object in our analysis is the (regularized) Gram matrix of the "tails" of the data, which we denote as

$$\boldsymbol{A}_k := \boldsymbol{Q}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top + \lambda \boldsymbol{I}_n. \tag{10}$$

Just as in (Tsigler and Bartlett, 2023) the main assumption under which our arguments work is that the condition number of the matrix $\boldsymbol{A}_k$ is bounded by some constant. Therefore, we introduce the following event.

Definition 2 *For any $k \in \{0, 1, \dots, p-1\}$ and $L \geq 1$, we denote by $\mathscr{A}_k(L)$ the event*

$$\mathscr{A}_k(L) := \left\{ \frac{1}{L}\Big(\lambda + \sum_{i>k}\lambda_i\Big) \leq \mu_n(\boldsymbol{A}_k) \leq \mu_1(\boldsymbol{A}_k) \leq L\Big(\lambda + \sum_{i>k}\lambda_i\Big) \right\}. \tag{11}$$

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

Throughout the paper we will also always impose assumptions of the form

$$\lambda + \sum_{i>k} \lambda_i > c \left(n\lambda_{k+1} + \sqrt{n \sum_{i>k} \lambda_i^2} \right), \tag{12}$$

where c is some large constant. It is closely related to saying that the event $\mathscr{A}_k$ holds with high probability as well as to the notions of effective ranks used by Bartlett et al. (2020) and Tsigler and Bartlett (2023). Indeed, consider the following lemma.

Lemma 3 (Lemma 16 from (Tsigler and Bartlett, 2023)) *Suppose that elements of $\boldsymbol{Z}$ are σ_x-sub-Gaussian and independent. There exists a constant c that only depends on σ_x such that, with probability at least $1 - ce^{-n/c}$,*

$$\begin{aligned}
\mu_1(\boldsymbol{A}_k) = \lambda + \mu_1(\boldsymbol{Q}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top) \leq & \lambda + \sum_{i>k} \lambda_i + c \left(n\lambda_{k+1} + \sqrt{n \sum_{i>k} \lambda_i^2} \right), \\
\mu_n(\boldsymbol{A}_k) = \lambda + \mu_n(\boldsymbol{Q}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top) \geq & \lambda + \sum_{i>k} \lambda_i - c \left(n\lambda_{k+1} + \sqrt{n \sum_{i>k} \lambda_i^2} \right).
\end{aligned}$$

Lemma 3 shows that if components of the data are independent and sub-Gaussian, then (12) 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 do not believe independence to be *necessary* for $\mathscr{A}_k(L)$ to hold. The same logic was followed in Tsigler and Bartlett (2023) (see their Section 4). Moreover, (Tsigler and Bartlett, 2023, Section 5) explains that much weaker conditions, such as sub-Gaussianity and some small-ball condition, are sufficient for the event $\mathscr{A}_k(L)$ to hold with high probability. In (Tsigler and Bartlett, 2023, Section 5.4), they even show that $\mathscr{A}_k(L)$ can hold with high probability for some heavy-tailed distributions.

3.2 Algebraic assumptions

Similarly to Tsigler and Bartlett (2023), our arguments for the main lower bound cleanly decompose into an algebraic and a probabilistic part. Hence, we do not 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 4 *For any $k \in \{0, 1, \ldots, p-1\}$ and $c_B > 0$, we let $\mathscr{B}_k(c_B)$ be the event on which all of the following hold:*

1. $\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \leq c_B n$ *and* $\mu_n(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \geq n/c_B$.
2. $\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \leq c_B n \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2$.

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. $tr(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \le c_B n k$.

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 5 *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 five bounds from the definition of $\mathscr{B}_k(c_B)$ hold with probability at least $1 - c_B e^{-n/c_B}$, where c_B only depends on σ_x. Bounds 1–4 were derived in the proof of (Tsigler and Bartlett, 2023, Theorem 5) (see the display in the middle of page 59). To transform their notation into ours, one needs to replace X by $\boldsymbol{Q}$, Σ by $\boldsymbol{\Sigma}$ and Z by $\boldsymbol{Z}$. The last statement follows directly from (Tsigler and Bartlett, 2023, Lemma 24), one just needs to plug in $\boldsymbol{\Sigma}^2$ instead of Σ. ■

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 by Tsigler and Bartlett (2023) (see their Section 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 Tikhomirov (2018) 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 inequality 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 (Abdalla and Zhivotovskiy, 2023, Theorem 2).

4. Main results

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

$$\sigma_\eta := 1/\sqrt{\ln\frac{3+\eta^{-1}}{2}}. \tag{13}$$

Next, for a given k, we define

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

Note that we do not track the dependence on k in the notation since we always introduce k before using it.

To explain how these quantities arise, let us consider the ridgeless case ($\lambda = 0$) and return to the geometric picture from Section 2. In Section 2.1 we saw that there are two main terms in the expression for $\boldsymbol{w}_{\mathrm{MNI}}^c$: $(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^{-1}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ and $\boldsymbol{\mu}_\perp$. Recall that our main goal is to evaluate their scalar products with $\boldsymbol{\mu}$ as well as norms in $\boldsymbol{\Sigma}$.

First of all, the quantity V corresponds to $\mathbb{E}_{\boldsymbol{y}}\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2$, where $\mathbb{E}_{\boldsymbol{y}}$ denotes expectation over the draw of $\boldsymbol{y}$. 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 quantity $\lozenge$ controls the magnitude of $\boldsymbol{\mu}^\top \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} = \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$.

The quantities B and M arise as $B \approx \|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}}^2$ and $M \approx \boldsymbol{\mu}^\top \boldsymbol{\mu}_\perp$. Interestingly, $\lozenge$, is just a rescaling of $\sqrt{B}$, so the bounds for $\|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}}$ and $|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|$ are very similar.

The quantity N is just a rescaling of M, the main reason for its introduction is that it is more convenient to work with expressions for $(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{w}_{\mathrm{ridge}}$ than $\boldsymbol{w}_{\mathrm{ridge}}$, and $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \approx n\Lambda^{-1}$. Our main bounds are given in terms of N, V and $\lozenge$. One interesting property of those quantities is that, similarly to Mach number in fluid dynamics, they are dimensionless: multiplying $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{1/2}$ by the same scalar does not change them. Therefore, these quantities can be seen as intrinsic descriptors that characterize the classification task, regardless of the arbitrary scale or units of the input data. Additionally, when it comes to the interpretation from Section 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 $\left\|(\boldsymbol{\Sigma} + \Lambda n^{-1}\boldsymbol{I}_p)^{-1}\boldsymbol{\mu}\right\|_{\boldsymbol{\Sigma}}^2$ (see Lemma 16 for a precise statement).

For the case of noisy labels discussed in Section 2.2, informally, V controls $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}^2$ and $\lozenge$ controls $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$. We say "informally" because our proofs do not 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.

By virtue of algebra, our results extend to ridge regression. Note that the only quantity that directly depends on the explicit regularization λ is Λ—the sum of λ and the energy of the tail of the covariance. Thus, as we already mentioned in Section 1.3, this tail imposes the implicit regularization on the learning problem. Explicit regularization only adds to

the implicit regularization from the data without qualitatively changing the results, and Λ represents the total effective regularization.

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

Lemma 6 (Relations between the main quantities) *Suppose that*

$$k \leq n \quad \text{and} \quad \Lambda > n\lambda_{k+1} \vee \sqrt{n \sum_{i>k} \lambda_i^2}. \tag{14}$$

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

As discussed in Section 1.5, V is nothing but the variance term from the regression literature (Bartlett et al., 2020; Tsigler and Bartlett, 2023). As in those papers, V is bounded by a constant, but can be arbitrarily small. Moreover, $\|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}}^2$ is exactly the bias term from (Bartlett et al., 2020; Tsigler and Bartlett, 2023). A reader familiar with the regression literature may notice, however, that our expressions for V and B are somewhat different from the main bounds in (Tsigler and Bartlett, 2023). This is because we choose a different presentation strategy: while Tsigler and Bartlett (2023) give bounds in a simpler form, they are only tight for the right choice of k. The way we formulate our bounds makes them tight for any choice of k under which the assumptions are satisfied at the cost of more involved expressions. We elaborate more on the differences in techniques from Tsigler and Bartlett (2023) in Appendix B.1, but a formal connection is established in the following proposition. The fact that the bounds are the same up to a constant multiplier for the right choice of k can be seen from Lemma 15.

Proposition 7

$$V \leq \frac{k}{n} + \Lambda^{-2} n \sum_{i>k} \lambda_i^2 \quad \text{and} \quad B \leq n^{-2}\Lambda^2 \|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}^2 + \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2.$$

Proof

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

■

4.1 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 8 (Main lower bound) *For any $c_B > 0$ and $L > 1$, 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 $S > 0$:

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} \geq c^{-1}N - ct\lozenge, \tag{15}$$

$$S\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}} \leq c\left([1 + N\sigma_\eta]\sqrt{V + t^2\Delta V} + \lozenge\sqrt{n}\right). \tag{16}$$

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

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

In Section 4.4 we explain how this result relates to the expressions we derived in Section 2.

The proof of Theorem 8 is the major part of the appendix, the outline can be found in Appendix B.1.

4.2 Upper bound

Obtaining an upper bound on $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}/\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ turns out to be more technically challenging. We elaborate on that in Appendix B.2, where we explain and justify which additional assumptions are needed to derive an upper bound. Due to these technical difficulties, we only provide the upper bound for the regime without label-flipping noise in the following theorem.

Theorem 9 (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 $\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} < 0$ with probability at least $c_y^{-1}(\mathbb{P}(\mathscr{A}_k(L)) - ce^{-n/c})_+$, where 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

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

The proof outline of Theorem 9 can be found in Appendix B.2.

4.3 Tight bound for a quantile

Theorems 8 and 9 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 a parameter t that controls the probability with which the bound holds. We do not expect those bounds to be sharp for all possible values of t, but we show that they are sharp when t is a constant.

Definition 10 *For any $\varepsilon \in (0,1)$, we denote by α_ε the ε-quantile of the distribution of $\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}/\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}$, i.e.,*

$$\alpha_\varepsilon := \inf\left\{ \alpha \in \mathbb{R} : \mathbb{P}\left(\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} < \alpha \right) > \varepsilon \right\}. \tag{17}$$

The following theorem, whose proof can be found in Appendix J, 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 11 (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,

$$\alpha_\varepsilon \leq c\frac{N}{\sqrt{V} + \sqrt{n}\lozenge}.$$

If additionally $N \geq a\lozenge$, then

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

4.4 Regimes of the lower bound

In this section, we relate the bound from Theorem 8 to the geometric story from Section 2. To do so, we set $\lambda = 0$ and plug expressions from Lemma 54 in Appendix I into the formulas from Section 2. For simplicity, we let t be a large constant and η a small constant, and only provide the derivation on an informal level using the $\approx$ and $\lesssim$ notation. The bounds of Lemma 54 then become

$$|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| \lesssim \lozenge, \quad |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}| \lesssim \lozenge, \quad |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}| \lesssim \lozenge,$$
$$\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \lesssim V, \quad \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \lesssim V, \quad \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \approx n\Lambda^{-1},$$
$$\boldsymbol{\mu}^\top \boldsymbol{\mu}_\perp \approx M, \quad \text{and} \quad \|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}} \lesssim \Lambda\lozenge/\sqrt{n},$$

where $\Delta\boldsymbol{y} := \hat{\boldsymbol{y}} - \boldsymbol{y}$. Note that instead of dealing with $\tilde{\boldsymbol{y}}$ as in Section 2.2, our bounds involve $\Delta\boldsymbol{y}$. Still, we speculate that for constant η, the terms involving $\tilde{\boldsymbol{y}}$ behave in the same way as those involving $\Delta\boldsymbol{y}$. Hence, for the rest of this section we additionally consider the following informal bounds:

$$|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}| \lesssim \lozenge, \quad \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \approx V, \quad \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}^2 \approx V, \quad \text{and} \quad \|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}} \approx \Lambda\lozenge/\sqrt{n}.$$

What is more, though the structure of our argument does not provide individual lower bounds on $\|\boldsymbol{\mu}_\perp\|_{\boldsymbol{\Sigma}}$ or $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2$, they are directly analogous to the bias and variance terms from regression, so that matching lower bounds were derived in (Bartlett et al., 2020; Tsigler and Bartlett, 2023).

As in Section 2, we separately consider noiseless and noisy cases. We also fix the covariance $\boldsymbol{\Sigma}$, the number of data points n, and the direction of $\boldsymbol{\mu}$, treating its magnitude $\|\boldsymbol{\mu}\|$ as a parameter. For each case, we analyze three scalar quantities for the various terms in the decomposition of $\boldsymbol{w}_{\text{MNI}}$: scalar product with $\boldsymbol{\mu}$, norm in $\boldsymbol{\Sigma}$, and regular Euclidean norm. Although the Euclidean norm does not appear directly in the classification accuracy bounds, it is still informative in our setting. Indeed, ridge regularization is defined through the Euclidean norm, so while $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}$ and $\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$ determine the behavior of the classification problem, $\|\boldsymbol{w}_{\text{MNI}}\|$ serves as an observable proxy—one that can be evaluated without knowledge of internal parameters such as $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$.

4.4.1 MNI without label-flipping noise

For $\eta = 0$, we obtained in Section 2.1 that

$$\tilde{\boldsymbol{w}}_{\text{MNI}}^c(\boldsymbol{\mu}) = \tilde{\boldsymbol{w}}_{\text{MNI}}^c(\boldsymbol{0}_p) + \boldsymbol{\mu}_\perp + \frac{\boldsymbol{\mu}^\top \tilde{\boldsymbol{w}}_{\text{MNI}}^c(\boldsymbol{0}_p)}{\|\tilde{\boldsymbol{w}}_{\text{MNI}}^c(\boldsymbol{0}_p)\|^2}\tilde{\boldsymbol{w}}_{\text{MNI}}^c(\boldsymbol{0}_p) \tag{18}$$
$$= \frac{\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}} + \underbrace{(\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu}}_{\boldsymbol{\mu}_\perp} + \frac{\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}. \tag{19}$$

As we have noted before, the first term in these expressions does not depend on $\boldsymbol{\mu}$ and has a symmetric distribution, thus acting as a "noise term." We also derived in Section 2.1, that the sum of the second and third terms is the projection of $\boldsymbol{\mu}$ on $\mathcal{Q}_A^\perp$, while the second term on its own is the projection of $\boldsymbol{\mu}$ on $\mathcal{Q}^\perp$. Hence, the second and third terms carry the signal, the second term being geometrically very similar to the sum of both of them.

Before plugging in the expressions from Lemma 54, we change the scaling of $\boldsymbol{w}_{\text{MNI}}$ to that used in Theorem 8, which provides bounds for $S\boldsymbol{w}_{\text{ridge}}$ instead of $\boldsymbol{w}_{\text{ridge}}$. In case of MNI, the scaling factor reduces to $S = \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\tilde{\boldsymbol{w}}^c_{\text{MNI}}\|^2$ and it is easy to see why multiplying by S is algebraically convenient: since $\boldsymbol{w}^c_{\text{MNI}} = \tilde{\boldsymbol{w}}^c_{\text{MNI}}/\|\tilde{\boldsymbol{w}}^c_{\text{MNI}}\|^2$, it allows us to work with $\tilde{\boldsymbol{w}}^c_{\text{MNI}}$ instead of $\boldsymbol{w}_{\text{MNI}}$. What is more, multiplying by $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ cancels out the denominator in (19). We can therefore write

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

We are now in a position to apply Lemma 54. Let us start with the term $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_\perp$, for which we obtain

$$\begin{aligned}
\boldsymbol{\mu}^\top(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_\perp) &= \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 \approx n\Lambda^{-1}M,\\
\|\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_\perp\|^2_{\boldsymbol{\Sigma}} &= (\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{\mu}_\perp\|^2_{\boldsymbol{\Sigma}} \lesssim n\lozenge^2, \quad \text{and}\\
\|\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_\perp\|^2 &= (\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{\mu}_\perp\|^2 \approx n^2\Lambda^{-2}M.
\end{aligned}$$

Recall that the direction along which the quantity $\boldsymbol{\mu}^\top\boldsymbol{w}/\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}$ is maximized is $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and denote the regularized version of this optimal vector as $\boldsymbol{w}_{\text{opt}}(\lambda) := (\boldsymbol{\Sigma} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\mu}$. In Section 5, we show that $n\Lambda^{-1}M \approx \boldsymbol{\mu}^\top\boldsymbol{w}_{\text{opt}}(\Lambda/n)$, $n\lozenge^2 \approx \|\boldsymbol{w}_{\text{opt}}(\Lambda/n)\|^2_{\boldsymbol{\Sigma}}$ (see Lemma 16 for a precise statement). Hence, in terms of both scalar product with $\boldsymbol{\mu}$ and norm in $\boldsymbol{\Sigma}$, the term $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_\perp$ behaves as $\boldsymbol{w}_{\text{opt}}(\Lambda/n)$. When it comes to the Euclidean norm, however, its norm is larger: $n^2\Lambda^{-2}M \approx n\Lambda^{-1}\boldsymbol{\mu}^\top\boldsymbol{w}_{\text{opt}}(\Lambda/n) > \|\boldsymbol{w}_{\text{opt}}(\Lambda/n)\|^2$.

When it comes to the term $\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$, our bounds become

$$\begin{aligned}
\boldsymbol{\mu}^\top(\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}) &= (\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 \lesssim \lozenge^2,\\
\|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|^2_{\boldsymbol{\Sigma}} &= (\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|^2_{\boldsymbol{\Sigma}} \lesssim \lozenge^2 V, \quad \text{and}\\
\|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|^2 &= (\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \lesssim n\Lambda^{-1}\lozenge^2.
\end{aligned}$$

Observe that these quantities are dominated by our bounds on $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_\perp$. Indeed, from Lemma 6 we get $\lozenge^2 \le n\lozenge^2 \le N = n\Lambda^{-1}M$, $\lozenge^2 V \le 2\lozenge^2 \le n\lozenge^2$ and $n\Lambda^{-1}\lozenge^2 \le \Lambda^{-1}N\sqrt{n\Delta V} \le \sqrt{3}\Lambda^{-1}N = \sqrt{3}n\Lambda^{-2}M \le \sqrt{3}n^2\Lambda^{-2}M$. This aligns with the geometric intuition that the projection of $\boldsymbol{\mu}$ on $\mathcal{Q}_A^\perp$ behaves similarly to the projection on $\mathcal{Q}^\perp$.

Finally, when it comes to the term $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$, we get

$$\begin{aligned}
|\boldsymbol{\mu}^\top(\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y})| &= \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \lesssim \lozenge,\\
\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|^2_{\boldsymbol{\Sigma}} &\lesssim V, \quad \text{and}\\
\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|^2 &= \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \approx n\Lambda^{-1}.
\end{aligned}$$

To obtain the bound from Theorem 8, we bound $S\boldsymbol{\mu}^\top\boldsymbol{w}^c_{\text{MNI}}$ from below and $S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$ from above, plugging in the results from above. Explicitly,

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

We then obtain, for a new constant c, that

$$S\boldsymbol{\mu}^\top\boldsymbol{w}^c_{\text{MNI}} \gtrsim N - c\lozenge \quad \text{and} \quad S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \lesssim \sqrt{V} + \sqrt{n}\lozenge. \tag{20}$$

Recall that N and $\Diamond$ have quadratic and linear dependences on $\boldsymbol{\mu}$ respectively and that V does not depend on $\boldsymbol{\mu}$. Hence, as the magnitude of $\boldsymbol{\mu}$ grows, there are two qualitative transitions in the bound: when N starts dominating $c\Diamond$ and when $\sqrt{n}\Diamond$ starts dominating $\sqrt{V}$. Interestingly, the first transition always happens before the second one: for small $\boldsymbol{\mu}$, the "noise term" $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}$ dominates $\boldsymbol{\mu}_\perp$ in terms of both scalar product with $\boldsymbol{\mu}$ and norm in $\boldsymbol{\Sigma}$. Then, $\boldsymbol{\mu}_\perp$ starts dominating in terms of scalar product, but is still behind in terms of the norm in $\boldsymbol{\Sigma}$. Finally, when $\boldsymbol{\mu}$ becomes sufficiently large, $\boldsymbol{\mu}_\perp$ dominates in terms of both, which corresponds to the "recovery of geometry" regime announced in Section 1.3. When it comes to the Euclidean norm, the transition between $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|^2 \approx n\Lambda^{-1}$ and $(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\|\boldsymbol{\mu}_\perp\|^2 \approx n\Lambda^{-1}N$ always happens after the transition in scalar product, but it may happen before or after the transition for the norm in $\boldsymbol{\Sigma}$.

The formal statement is given by the following proposition:

Proposition 12 (Transition order, noiseless case) *Assume the conditions of Lemma 6. Fix the direction of $\boldsymbol{\mu}$ and treat its magnitude as a parameter. For any scalar quantities a, b define by $\overset{b}{\nearrow\!\!\!/}_a$ the smallest magnitude of $\boldsymbol{\mu}$ for which $b \geq a$.*

For any C, for any $n > 8C^2$, one of the following two chains of inequalities holds.

- *Case 1:*
$$\overset{N}{\nearrow\!\!\!/}_{C\Diamond} \leq \overset{\sqrt{n}\Diamond}{\nearrow\!\!\!/}_{2C\sqrt{V}} \leq \overset{N}{\nearrow\!\!\!/}_{2C}\,.$$
- *Case 2:*
$$\overset{N}{\nearrow\!\!\!/}_{C\Diamond} \leq \overset{N}{\nearrow\!\!\!/}_{2C} \leq \overset{\sqrt{n}\Diamond}{\nearrow\!\!\!/}_{2C\sqrt{V}}\,.$$

Moreover, both cases are possible depending on $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$.

Proof This is a direct consequence of Proposition 13 proved in Appendix E. ■

4.4.2 MNI with label-flipping noise

In this section, we tie the geometric story of Section 2.2 to Theorem 8, investigating the different regimes that arise in the presence of label-flipping noise. To do so, we rely on the representation in (9) of the MNI solution for the noisy problem in terms of the one for the clean one, namely,

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

As in Section 4.4.1, we rescale this quantity by $S = \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\tilde{\boldsymbol{w}}^c_{\text{MNI}}\|^2$ to use the bounds from Theorem 8. Explicitly, the remainder of this section deals with

$$S\boldsymbol{w}_{\text{MNI}} = S\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}} + \xi S\boldsymbol{w}^c_{\text{MNI}} - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}} S\boldsymbol{w}^c_{\text{MNI}}. \tag{21}$$

Since $\xi \approx 1$, the second term is essentially the same vector we worked with in Section 4.4.1 given that $S\boldsymbol{w}^c_{\text{MNI}} = \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\tilde{\boldsymbol{w}}^c_{\text{MNI}}$. The first term is now the "noise component" $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ scaled by the squared norm of $\tilde{\boldsymbol{w}}^c_{\text{MNI}}$ and $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \approx n\Lambda^{-1}$. As in Section 4.4.1, we are going to look at $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}$, $\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$, and $\|\boldsymbol{w}_{\text{MNI}}\|$.

Let us start with the Euclidean norm $\|\boldsymbol{w}_{\text{MNI}}\|$. Recall that $\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ is orthogonal to $\boldsymbol{w}^c_{\text{MNI}}$. From Section 4.4.1, since $\boldsymbol{\mu}_\perp$ is orthogonal to $\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{y}$, we have $\|S\boldsymbol{w}^c_{\text{MNI}}\|^2 \approx n\Lambda^{-1}(1+N)$. At the same time, $\|S\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|^2 \approx (1+N)^2 n\Lambda^{-1}$, $\xi \approx 1$, and $|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}| \lesssim \lozenge$. Since $N \gtrsim \lozenge^2$, the norm of the "noise component" $S\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ is always at least a constant fraction of the norms of the remaining terms. That is, in Euclidean geometry, the noisy solution $\boldsymbol{w}_{\text{MNI}}$ looks substantially different than the clean solution $\boldsymbol{w}^c_{\text{MNI}}$: it picks up an orthogonal term, whose norm is larger (up to a constant factor) than the norm of the clean solution. Moreover, the component in the direction of the clean solution points in the opposite direction with probability close to 0.5 for large enough $\boldsymbol{\mu}$. Using the notation of Proposition 12, that transition happens at $\nearrow^{1}_{\lozenge}$. As we show in Proposition 13, this transition happens after all the transitions in the main bound.

Though the Euclidean geometry of $\boldsymbol{w}_{\text{MNI}}$ and $\boldsymbol{w}^c_{\text{MNI}}$ look very different, when it comes to $\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}$ and $\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$, the bounds from Theorem 8 for their inner products with $\boldsymbol{\mu}$ and their norms with respect to $\boldsymbol{\Sigma}$ turn out to be very similar. Unfortunately, we do not have an intuitive explanation as to why $S\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}}$ resembles $S\boldsymbol{\mu}^\top \boldsymbol{w}^c_{\text{MNI}}$. As we mentioned in Section 2.2, algebraic cancellations cause the formulae for the scalar product in the noisy and noiseless cases to be similar, namely,

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

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. This similarity implies that the conditions under which the bound becomes positive are the same: using the notation of Proposition 12, this transition happens at $\nearrow^{C\lozenge}_{N}$.

When it comes to $S\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$, the story is more clear. The contribution of the third term from (21) turns out to be dominated by the first two terms. To see this, note that

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

We already have the upper bound on $S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}}$ from Section 4.4.1, namely,

$$S\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \lesssim \sqrt{V} + \lozenge\sqrt{n}.$$

The other bounds follow from

$$S = (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 \lesssim (1 + \lozenge)^2 + N \approx 1 + N,$$
$$\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} \lesssim \sqrt{V}, \quad \xi \approx 1, \quad \text{and} \quad |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}| \lesssim \lozenge.$$

Combine these results yields

$$\begin{aligned} S\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} &\lesssim (1 + N)\sqrt{V}, \\ \xi S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} &\lesssim \sqrt{V} + \sqrt{n}\lozenge, \\ |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}|S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} &\lesssim \lozenge\sqrt{V} + \sqrt{n}\lozenge^2. \end{aligned}$$

Using Lemma 6, we see that $N\sqrt{V} \geq N\sqrt{4\Delta V} \geq N\sqrt{12/n} \geq \sqrt{12n}\Diamond^2$. Moreover, $\sqrt{n}\Diamond$ dominates $\Diamond\sqrt{V}$. That is, the bound on the last term is dominated (up to a constant factor) by the bounds on the first two:

$$S\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \lesssim (1+N)\sqrt{V} + \sqrt{V} + \sqrt{n}\Diamond \approx N\sqrt{V} + \sqrt{V} + \sqrt{n}\Diamond.$$

Let us now focus on the geometry behind the first two terms in (21): $S\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ and $\xi S\boldsymbol{w}^c_{\text{MNI}}$. Firstly, when $\boldsymbol{\mu}$ is small, then $S \approx 1$, $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} \approx \sqrt{V}$, and $\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} = \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}} \approx \sqrt{V}$. That is, for small $\boldsymbol{\mu}$, both terms are of the same order, so $S\boldsymbol{w}_{\text{MNI}}$ and $S\boldsymbol{w}^c_{\text{MNI}}$ display the same behavior in terms of their scalar product with $\boldsymbol{\mu}$ and their norm in $\boldsymbol{\Sigma}$. Thus, both noisy and noiseless bounds go through the same initial regime $N/\sqrt{V}$.

Yet, $S\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}$ grows quadratically with $\boldsymbol{\mu}$, while $S\boldsymbol{w}^c_{\text{MNI}} = \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\tilde{\boldsymbol{w}}^c_{\text{MNI}}$ grows linearly. Thus, as $\boldsymbol{\mu}$ becomes very large, the second term becomes vanishingly small and $S\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \approx S\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} \approx N\sqrt{V}$. The bound then becomes $1/\sqrt{V}$.

What happens between these regimes, i.e., for moderate magnitudes of $\boldsymbol{\mu}$, depends on the relationship between $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. Indeed, we can write

$$\begin{aligned} S\|\boldsymbol{w}^c_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \approx& \sqrt{V} + \sqrt{n}\Diamond, \\ S\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}} \approx& (N+1)\sqrt{V}. \end{aligned}$$

The bound is the ratio of these quantities, and both numerator and denominator are increasing, so there is a race between the growth in Euclidean geometry and in the norm in $\boldsymbol{\Sigma}$.

The "recovery of geometry" regime appears if for a certain magnitude of $\boldsymbol{\mu}$ we have $\sqrt{n}\Diamond \gtrsim \max(\sqrt{V}, N\sqrt{V})$. Since V does not scale with $\boldsymbol{\mu}$, $\sqrt{n}\Diamond$ scales linearly, and $N\sqrt{V}$ scales quadratically, the condition for such magnitude of $\boldsymbol{\mu}$ to exist is $NV \lesssim n\Diamond^2$. Calling once again upon Lemma 16, we can rewrite this condition as

$$\sum_i \frac{\mu_i^2 V}{\lambda_i + \Lambda/n} \lesssim \sum_i \frac{\lambda_i \mu_i^2}{(\lambda_i + \Lambda/n)^2}.$$

Thus, a sufficient condition is for $\boldsymbol{\mu}$ to be concentrated on the coordinates in which $V \lesssim \lambda_i/(\lambda_i + \Lambda/n)$. Intuitively, the recovery of geometry appears if $\boldsymbol{\mu}$ has most of its mass in the first several coordinates.

The following proposition, whose proof can be found in Appendix E, shows that there are only two cases for the order in which transitions between domination of different terms happen. Figure 1 provides a diagram with transitions of various scalar quantities for those two cases.

Proposition 13 (Transition order, noisy case) *Assume the conditions of Lemma 6. Fix the direction of $\boldsymbol{\mu}$ and treat its magnitude as a parameter. For any scalar quantities a, b define by ${}^{b}\!\nearrow_{a}$ the smallest magnitude of $\boldsymbol{\mu}$ for which $b \geq a$.*

For any C, for any $n > 8C^2$, one of the following two chains of inequalities holds.

- *Case 1:*

$${}^{N}\!\nearrow_{C\Diamond} \leq {}^{\sqrt{n}\Diamond}\!\nearrow_{2C\sqrt{V}} \leq {}^{N}\!\nearrow_{2C} \leq {}^{N\sqrt{V}}\!\nearrow_{\sqrt{n}\Diamond} \leq {}^{\Diamond}\!\nearrow_{1}.$$

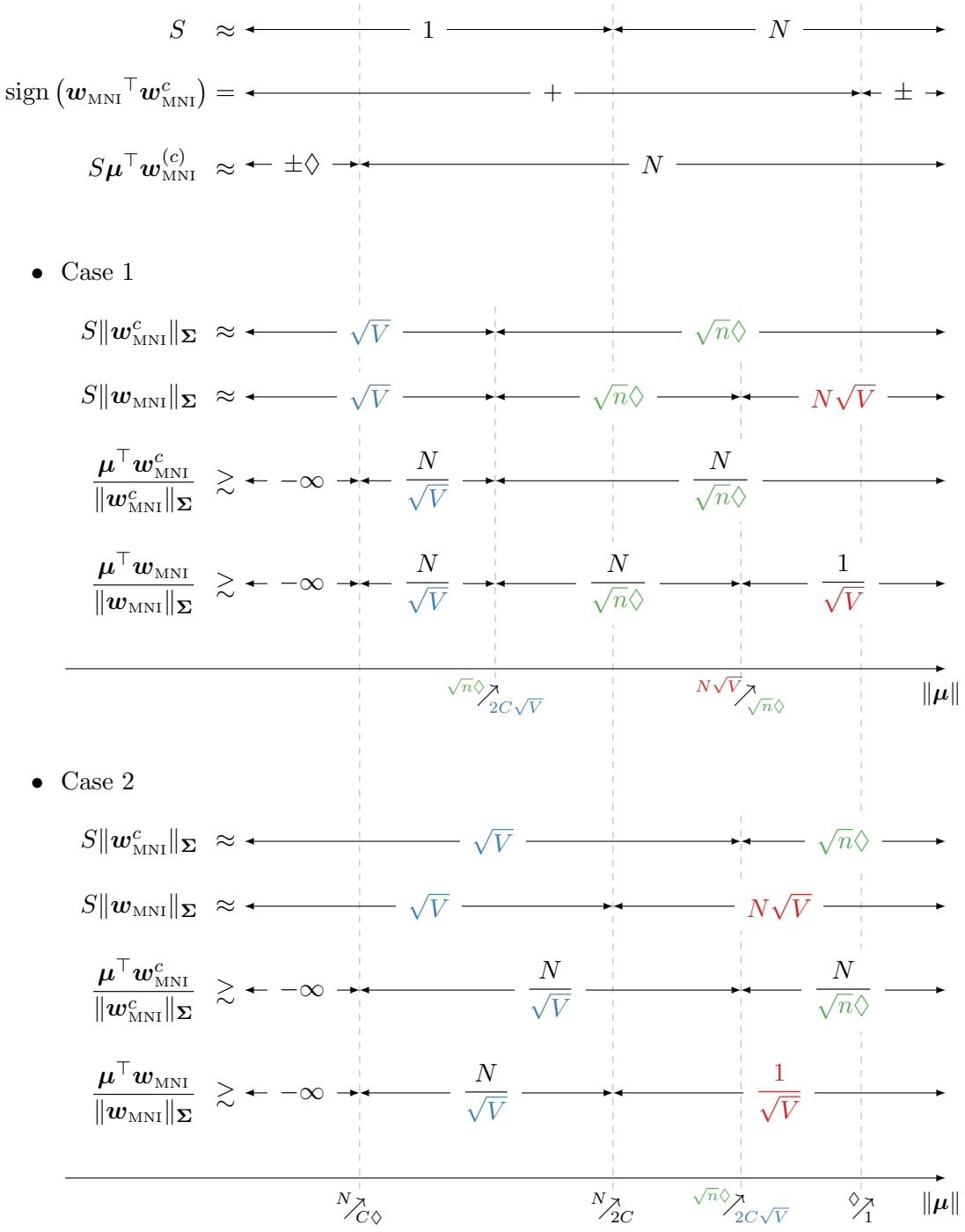


Figure 1: Main bounds on the quantities of interest (up to a constant factor) as a function of $\|\boldsymbol{\mu}\|$. We write $\boldsymbol{w}^{(c)}_{\mathrm{MNI}}$ to denote both $\boldsymbol{w}_{\mathrm{MNI}}$ and $\boldsymbol{w}^{c}_{\mathrm{MNI}}$.

- *Case 2:*

$$\frac{N}{C\Diamond} \le \frac{N\sqrt{V}}{\sqrt{n}\Diamond} \le \frac{N}{2C} \le \frac{\sqrt{n}\Diamond}{2C\sqrt{V}} \le \frac{\Diamond}{1}.$$

Moreover, both cases are possible depending on $\boldsymbol{\Sigma}$ *and* $\boldsymbol{\mu}$.

4.5 Benign overfitting

Expanding on the informal observations from the end of last section, one interesting phenomenon that Theorem 8 implies is benign overfitting: the misclassification error can be arbitrarily close to zero even if $\boldsymbol{w}_{\text{MNI}}$ interpolates training labels with a small constant level of label-flipping noise. One can see that it happens for $t = o(\sqrt{n})$, $V = o(1)$, and $N = \omega(\sqrt{V} + t/\sqrt{n} + \Diamond\sqrt{n})$ (which has the form of an SNR condition). In order to simplify the presentation, we formulate a rigorous corollary for the Gaussian distribution.

Corollary 14 *Suppose the rows of matrix* $\boldsymbol{Q}$ *come from i.i.d. samples from a Gaussian distribution. Take* $\lambda = 0$ *(that is, consider the MNI solution). There exists a large absolute constant* c *such that, for any* $C > 1$*, if*

- *the noise is bounded by a constant:* $$\eta < c^{-1}, \tag{22}$$
- *the "spiked part" of the covariance has low dimension:* $$k < n/(cC^2), \tag{23}$$
- *the "tail" of the covariance has high effective rank:* $$\sum_{i>k} \lambda_i > cn\lambda_{k+1} \vee cC\sqrt{n\sum_{i>k}\lambda_i^2}, \tag{24}$$
- *the scale of* $\boldsymbol{\mu}$ *is large enough:* $$N \ge 1 + cC\left(\sqrt{V} + \Diamond\sqrt{n}\right), \tag{25}$$

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

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{MNI}}{\|\boldsymbol{w}_{MNI}\|_{\boldsymbol{\Sigma}}} \ge C. \tag{26}$$

Proof First of all, due to (24), if c is large enough, by Lemma 3, 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 (23) and Lemma 5, 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 8, 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 $1 - c_1 e^{-t^2/2}$,

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}} \ge c_1^{-1} N - c_1 t\Diamond, \tag{27}$$

$$S\|\boldsymbol{w}_{\text{MNI}}\|_{\boldsymbol{\Sigma}} \le c_1\left(\left[1 + N\sigma_\eta\right]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n}\right). \tag{28}$$

Since $t < \sqrt{n}/c_1$, if c is large enough, due to (25) and (27), there exists an absolute constant c_2 such that $S\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{MNI}} \ge c_2^{-1} N$.

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

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

For simplicity, we use Lemma 6 to bound $\Delta V < 3/n$. We see that to achieve the bound in (26), we need to show the following two conditions:

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

Recall that $\eta < c^{-1}$, so that $\sigma_\eta < 1$. The two conditions above can therefore be achieved by first taking $t^2 = n/(c_4C^2)$ for a large enough absolute constant c_4. Then, 25 implies the first condition and (23)–(24) imply the second due to Proposition 7. ■

We believe these sufficient conditions for benign overfitting in classification to be novel. For example, in a recent paper, Wang and Thrampoulidis (2021) obtain similar conditions only for the case of isotropic data, i.e., $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ (see their Theorem 7).

As we explained in Section 4.4.2, we only provide a tight bound for the case without label-flipping noise, and thus, can only state sufficient conditions for benign overfitting. In view of our derivations in Sections 2.2 and 4.4.2, we expect the necessary conditions to be very similar, but leave the exact characterization for future work.

5. Effect of regularization

In this section we discuss the effects of the regularization on the accuracy of the learned classifier in the noiseless setting (i.e., $\eta = 0$). We will touch on the noisy setting in Section 5.5, 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 11, 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, our bounds are closely related to bounds from the earlier work on regression (Tsigler and Bartlett, 2023) but have a somewhat different form. In the next two lemmas we show how our quantities of interest could be alternatively defined to have a similar form to the quantities from that paper.

The following lemma gives a form of the bounds using the notion of k^* introduced by Bartlett et al. (2020). This form corresponds to the form of the main results obtained by Tsigler and Bartlett (2023).

Lemma 15 (Bounds via k^*) *Suppose that*

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

Define

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

Then

$$2N_* \geq N \geq N_*/2, \quad 2\lozenge_* \geq \lozenge \geq \lozenge_*/2, \quad 4V_* \geq V \geq V_*/4, \quad \Lambda_* \geq \Lambda \geq \Lambda_*/2.$$

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 7 of (Tsigler and Bartlett, 2023).

Lemma 16 (Alternative form of the bounds) *Suppose that $k < n$ and $\Lambda > n\lambda_{k+1}$. Denote*

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

Then

$$N \geq N_a \geq N/2, \quad V \geq V_a \geq V/4, \quad \lozenge^2 \geq \lozenge_a{}^2 \geq \lozenge^2/4.$$

Note that this form of the results already appeared in our discussion in Section 1.3.

Now we are in a position to return to studying the effect of regularization. To track changing values of the regularization parameter λ, for the rest of this section we add it explicitly to the notation, i.e., in this section we will write $\alpha_\varepsilon(\lambda), \Lambda(\lambda), N(\lambda), V(\lambda), \lozenge(\lambda), \mathscr{A}_k(L,\lambda)$ 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.

5.1 Increasing regularization never helps in the noiseless case

Due to Theorem 11, 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 16, 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

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

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

Lemma 17 *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 $\sqrt{n}\Diamond$ 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 K.

Lemma 18 (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 $\lambda' > \lambda$*

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

Combining this with Theorem 11 gives the following.

Corollary 19 *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 $n > c$, $k < n/c$, $\mathbb{P}(\mathscr{A}_k(L, \lambda)) > 1 - \delta$, $N(\lambda) \ge a\Diamond(\lambda)$, 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 $\boldsymbol{Q}_{0:k}$.

For every $\lambda' \ge \lambda$

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

Proof Take a, δ, ε the same as in Theorem 11. Denote the constant c from that theorem as c_1. Note that $\mathbb{P}(\mathscr{A}_k(L, \lambda')) \ge \mathbb{P}(\mathscr{A}_k(L, \lambda)) > 1 - \delta$, which means that Theorem 11 applies for all values of the regularization parameter $\lambda' > \lambda$. Thus,

$$\alpha_\varepsilon(\lambda') \le c_1 \frac{N(\lambda')}{\sqrt{V(\lambda')} + \sqrt{n}\Diamond(\lambda')}, \quad \alpha_\varepsilon(\lambda) \ge c_1^{-1} \frac{N(\lambda)}{\sqrt{V(\lambda)} + \sqrt{n}\Diamond(\lambda)}.$$

Combining it with Lemma 18 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".

5.2 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 in the tail 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 K.

Corollary 20 (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

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

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 Wang and Thrampoulidis (2021), 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 Wang and Thrampoulidis (2021) 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 6.2.

5.3 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 15 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:

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

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

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

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 $n\Lambda_*^{-1}\|\boldsymbol{\mu}_{k^*:\infty}\|^2/\|\boldsymbol{\mu}_{0:k^*}\|_{\boldsymbol{\Sigma}^{-1}_{0:k^*}}$ 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 $\Lambda_* \leq n\lambda_{k^*}$. Putting it together with $\Lambda_* \geq n\lambda_{k^*+1}$ 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 K.

Lemma 21 *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,*

$$\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \le C^{-1}n^{-1}\Lambda(\lambda)\|\boldsymbol{\mu}_{k:\infty}\|^2.$$

4. *Take*[3] $\boldsymbol{\mu}_{0:k}$ *which balances* $\boldsymbol{\mu}_{k:\infty}$ *in the following sense:*

$$nC^{-1}\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2 \ge \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \ge n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}. \tag{29}$$

5. *Scale* $\boldsymbol{\mu}$ *up*[4] *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 K, shows a particular example when the optimal regularization is negative:

Corollary 22 *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.*

5.4 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 $(\boldsymbol{w}_{\text{ridge}}(\lambda))_{0:k^*} \propto \boldsymbol{\Sigma}^{-1}_{0:k^*}\boldsymbol{\mu}_{0:k^*}$ 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 $\|\boldsymbol{\mu}_{0:k^*}\|_{\boldsymbol{\Sigma}^{-1}_{0:k^*}}$ 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 K:

Lemma 23 *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:*

3. Note that such $\boldsymbol{\mu}_{0:k}$ exists because of how we chose $\boldsymbol{\mu}_{k:\infty}$.
4. Note that the previous conditions were homogeneous in $\boldsymbol{\mu}$, so multiplying it by a scalar does not break them.

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*

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

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*

$$\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}} \geq C\|\boldsymbol{\mu}_{0:k}\|^2. \tag{30}$$

4. *Scale $\boldsymbol{\mu}$ up if needed, so that*

$$n\Diamond^2(\lambda) \geq V(\lambda) \quad \text{and } N(\lambda) \geq a\Diamond(\lambda).$$

Then for any λ' such that $\Lambda(\lambda') \geq n\lambda_1$

$$\alpha_\varepsilon(\lambda) \geq \frac{C}{c}\alpha_\varepsilon(\lambda').$$

A natural question is when one can choose such $\boldsymbol{\mu}_{0:k}$ that Equation (30) is satisfied. The answer is given by the following.

Lemma 24 *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: $\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}$ is the weighted average of $\{\lambda_i^{-1}\}_{i=1}^k$ with weights $\{\mu_i^2\}$, while $\|\boldsymbol{\mu}_{0:k}\|^{-2}_{\boldsymbol{\Sigma}_{0:k}}$ is inverse of the weighted average of $\{\lambda_i\}_{i=1}^k$. Thus, for the convex function $f(x) = 1/x$ we can write

$$\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}}\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}} = \frac{\sum_{i=1}^k \mu_i^2 f(\lambda_i)}{f\left(\sum_{i=1}^k \mu_i^2\lambda_i\right)}.$$

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:

$$\frac{\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}}\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}}{\|\boldsymbol{\mu}_{0:k}\|^4} = \left(\lambda_k + (\lambda_1-\lambda_k)\mu_1^2\right)\left(\lambda_k^{-1} + (\lambda_1^{-1}-\lambda_k^{-1})\mu_1^2\right).$$

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 K, shows another example when the optimal regularization is negative:

Corollary 25 *There exist absolute constants $b > c$ such that the following holds. Take $p > bn$, and $b \le k < n/b$. Consider the following classification problem with Gaussian data (in dimension p) and no label-flipping noise ($\eta = 0$):*

$$\lambda_i = \begin{cases} k^{-4i/k}, & i \le k, \\ \frac{cn}{pk^4}, & i > k. \end{cases}, \quad \mu_i = \begin{cases} \frac{b\ln(k)}{k^5}\left(\frac{k}{n} + \frac{n}{p}\right), & i \le k, \\ 0, & i > k. \end{cases}$$

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

5.5 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 8. 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 two 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 4.4, 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.

6. Detailed comparison with previous results

In this section, we compare our results with the results of four recent papers: (Cao et al., 2021), (Wang and Thrampoulidis, 2021), (Chatterji and Long, 2021), and (Muthukumar et al., 2021). Sections 6.1 and 6.2 show that our results generalize the results of the first two of those papers. In Section 6.3 we discuss the relation between our results and those of Chatterji and Long (2021), and show that their bound is weaker then ours for the case of Gaussian data. Finally, the paper of Muthukumar et al. (2021) considers a slightly different model. In Section 6.4 we explain how this model is related to ours, and show that some of their conclusions can be recovered from our analysis too, even though our results do not strictly generalize theirs.

Before we begin the comparisons, 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

$$\boldsymbol{w}_{\text{MM}} = \text{argmin}_{\boldsymbol{w}\in\mathbb{R}^p} \|\boldsymbol{w}\| \text{ s.t. } \boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{X}\boldsymbol{w} \geq \boldsymbol{1}_n, \tag{31}$$

where $\boldsymbol{D}_{\hat{\boldsymbol{y}}} = \text{diag}(\hat{\boldsymbol{y}})$. A standard argument with Lagrange multipliers shows that the solution to the optimization problem (31) 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 (31) 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 Muthukumar et al. (2021); Cao et al. (2021); Wang and Thrampoulidis (2021) 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., $\boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{X}\boldsymbol{w} = \boldsymbol{1}_n$, 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 Chatterji and Long (2021), they don't explicitly rely on support proliferation to study MM, but, as we explain in Section 6.3, 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 $k = 0$.

6.1 Comparison with (Cao et al., 2021)

Cao et al. (2021) study the same data generating model as ours. Their main result, reformulated in our notation, is given by the following theorem.

Theorem 26 (Theorem 3.1 and Proposition 4.1 from (Cao et al., 2021)) *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*

$$tr(\boldsymbol{\Sigma}) \geq C \max\left\{ n^{3/2}\|\boldsymbol{\Sigma}\|, n\|\boldsymbol{\Sigma}\|_F, n\sqrt{\log(n)}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \right\}, \tag{32}$$

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*

$$\frac{(\boldsymbol{\mu}^\top \boldsymbol{w}_{MNI})^2}{\|\boldsymbol{w}_{MNI}\|_{\boldsymbol{\Sigma}}^2} \geq C' \frac{n\|\boldsymbol{\mu}\|^4}{n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 + \|\boldsymbol{\Sigma}\|_F^2 + n\|\boldsymbol{\Sigma}\|^2}. \tag{33}$$

Note that if we take $k = 0$ and $\lambda = 0$, the assumption imposed in Equation (32) 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 3 and 5. Therefore, our Theorem 8 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 L, shows that our bound is at least as good as the bound from Cao et al. (2021).

Proposition 27 *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* $t < \sqrt{n}$,

$$\frac{N - ct\lozenge}{\sqrt{V + t^2\Delta V + \sqrt{n}\lozenge}} \geq \frac{1}{4} \frac{n\|\boldsymbol{\mu}\|^2}{n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} + \sqrt{n}\|\boldsymbol{\Sigma}\|_F + n\|\boldsymbol{\Sigma}\|}. \tag{34}$$

Note that the resulting bound does not depend on λ. We have already observed that in Corollary 20: 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 $k = 0$, Cao et al. (2021) did not observe the effect of "recovering the geometry."

6.2 Comparison with (Wang and Thrampoulidis, 2021)

The next paper we compare ours with is (Wang and Thrampoulidis, 2021). 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 28 (Theorem 3 from (Wang and Thrampoulidis, 2021)) *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*

$$n\lambda_1 < b\sum_i \lambda_i. \tag{35}$$

Take $\lambda \geq 0$. *Assume that* $\|\boldsymbol{\mu}\|^2 \geq a\left(n\Lambda^{-1}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}\right)$. *Then with probability at least* $1 - e^{-n^2/c}$

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} \geq c^{-1} \frac{\|\boldsymbol{\mu}\|^2 - a\left(n\Lambda^{-1}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}\right)}{(1 \vee n\Lambda^{-1}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}})\sqrt{\sum_i \lambda_i^2} + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}. \tag{36}$$

We see that their bound is at most within a constant factor of

$$\frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\Sigma}\|_F + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \frac{n\|\boldsymbol{\Sigma}\|_F}{\lambda + \operatorname{tr}(\boldsymbol{\Sigma})}}.$$

Comparing to our bound from Equation 34, we see that the bound from (Wang and Thrampoulidis, 2021) 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 (34) (after dividing both numerator and denominator by n). Moreover, it picks up an additional term $\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \frac{n\|\boldsymbol{\Sigma}\|_F}{\lambda + \operatorname{tr}(\boldsymbol{\Sigma})}$ in the denominator. Thus, the bound from Theorem 28 is worse than the one from Equation (34). Note that, just as in the previous section, Equation (35) implies that our Theorem 8 is applicable with high probability. That is, Proposition 27 shows that our result generalizes the result for balanced ensembles from Wang and Thrampoulidis (2021).

When it comes to the bi-level ensemble, the result of Wang and Thrampoulidis (2021) translated to our notation is given by the following theorem.

Theorem 29 (Theorem 5 from (Wang and Thrampoulidis, 2021)) *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 $\|\boldsymbol{\mu}\|^2 \geq a\left(n\Lambda^{-1}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}\right)$. Assume that $k = 1$, $\lambda \geq 0$ and*

$$bn\lambda_1 > \sum_{i>1} \lambda_i \quad \text{and} \quad bn\lambda_2 < \sum_{i>2} \lambda_i. \tag{37}$$

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 $1 - e^{-n/c}$,

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge}}{\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}}} \geq c^{-1} \frac{\|\boldsymbol{\mu}\|^2(1 - cn\Lambda^{-1}\lambda_j) - c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}{A + B + \lambda_j + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}. \tag{38}$$

Note that for $k = 1$ the second part of Equation (37) 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 3 and 5. Therefore, our Theorem 8 is applicable just as it was in Section 6.1. So, the following proposition, whose proof can be found in Appendix L, shows that our bound generalizes the bound for bi-level ensembles from (Wang and Thrampoulidis, 2021).

Proposition 30 *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 29. Then for $t \leq \sqrt{n}$*

$$\frac{N - ct\lozenge}{\sqrt{V + t^2\Delta V} + \sqrt{n}\lozenge} \geq \frac{1}{6} \frac{\|\boldsymbol{\mu}\|^2}{A + B + \lambda_j + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}.$$

Overall, we see that the bounds from Wang and Thrampoulidis (2021) 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."

6.3 Comparison with (Chatterji and Long, 2021)

Chatterji and Long (2021) 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 Chatterji and Long (2021) is not straightforward for two reasons. First, they consider the MM solution, while we consider the ridge and MNI solutions. Second, Chatterji and Long (2021) 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 3 and 5).

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 Frei et al. (2023), for max-margin to coincide with MNI, it suffices for the training data to be 'nearly-orthogonal' in the sense that for every training sample $(\boldsymbol{x}_k, y_k)$ it holds that $\|\boldsymbol{x}_k\|^2 \gg n \max_{i,j} \frac{\|\boldsymbol{x}_i\|^2}{\|\boldsymbol{x}_j\|^2} \max_{i \neq j} |\boldsymbol{x}_i^\top \boldsymbol{x}_j|$. 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 Chatterji and Long (2021). When translated into our notation, that result is given by the following.

Theorem 31 (Theorem 1 from (Chatterji and Long, 2021), 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 $\boldsymbol{X}, \hat{\boldsymbol{y}}$

$$\mathbb{P}_{(\boldsymbol{x}, \hat{y})}(\hat{y} \boldsymbol{x}^\top \boldsymbol{w}_{MM} < 0) \leq \eta + \exp\left(-b \frac{\|\boldsymbol{\mu}\|^4}{p}\right), \tag{39}$$

where $(\boldsymbol{x}, \hat{y})$ is a new data point from the data distribution with label-flipping noise, and $\mathbb{P}_{(\boldsymbol{x}, \hat{y})}$ denotes probability over the draw of that point.

Note that assumptions of Theorem 31 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 3 and 5, 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 8 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 31 we should take $t = \sqrt{2\log(1/\delta)}$. Finally, in Section 1.1 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 $\Phi(-z) \le e^{-z^2/2}$ for every $z > 0$, to recover the result of Theorem 31 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 L, shows that our result is stronger than Theorem 31.

Proposition 32 *Assume that $\lambda_i \le 1$ for any i and $\sum_{i=1}^p \lambda_i \ge \kappa p$ for some constant $\kappa \in (0, 1]$. Take $k = 0$, $\lambda = 0$ and some $c > 1$. Suppose additionally that $\kappa p/n \ge \|\boldsymbol{\mu}\|^2 \ge (2ct)^2/(\kappa^2 n)$, and $t^2 < n\kappa$.*

Then

$$\frac{N - ct\lozenge}{[1 + N\sigma_\eta]\sqrt{V + t^2\Delta V} + \lozenge\sqrt{n}} \ge \frac{1}{10}\frac{\|\boldsymbol{\mu}\|^2\sqrt{n\kappa}}{\sqrt{p}}.$$

That is, our lower bound picks up an additional factor of $\sqrt{n}$ compared to the bound from Theorem 31.

6.4 Comparison with (Muthukumar et al., 2021)

Apart from the data generating model considered in this paper, 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

$$\mathbb{P}\left(\operatorname{sign}(\boldsymbol{z}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}) \ne \operatorname{sign}(\boldsymbol{z}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha})\right) = \frac{\angle(\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha})}{\pi} = \frac{1}{\pi}\arccos\left(\frac{\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}\boldsymbol{w}}{\|\boldsymbol{\alpha}\|_{\boldsymbol{\Sigma}}\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}}\right),$$

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 paper: 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 also 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, the results in our paper 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 Muthukumar et al. (2021), who consider such a model. The main result of Muthukumar et al. (2021) 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

$$\lambda_i = \begin{cases} n^{s-q-r}, & \text{for } i \le n^r, \\ (1-n^{-q})/(1-n^{r-s}) & \text{for } i > n^r. \end{cases} \tag{40}$$

Finally, Muthukumar et al. (2021) consider $\boldsymbol{\alpha} = \boldsymbol{e}_1$ (note that, similarly to (Wang and Thrampoulidis, 2021), 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 $\boldsymbol{m} = \sqrt{\frac{2}{\pi\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{\alpha}}}\boldsymbol{\Sigma}\boldsymbol{\alpha} = \sqrt{\frac{2\lambda_1}{\pi}}\boldsymbol{e}_1$.

Muthukumar et al. (2021) 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 33 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 Muthukumar et al. (2021) precisely.

Proposition 33 *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 (40). 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:

$$\frac{N - ct\lozenge}{\sqrt{V + t^2\Delta V} + \sqrt{n}\lozenge} = (1+o_n(1))\frac{N}{\sqrt{V}+\sqrt{n}\lozenge} = \begin{cases} o_n(1), & 2q+2r-1-s > 0, \\ \frac{1+o_n(1)}{\sqrt{2\pi}} & 2q+2r-1-s = 0, \\ \sqrt{\frac{2}{\pi}} + o_n(1) & 2q+2r-1-s < 0. \end{cases}$$

Here we use $o_n(1)$ to denote quantities that converge to zero as n goes to infinity.

7. Conclusions and further directions

In this paper 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 literature.

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 a 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 from the regression literature (Bartlett et al., 2020), and it was shown there that high effective rank of the tail is necessary for benign overfitting to occur. Proving the necessity of this regime is another direction of future work.

Finally, even though we use a very similar regime to that considered in the regression literature, 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.

Acknowledgments

We gratefully acknowledge the support of the NSF for FODSI through grant DMS-2023505, of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639, and of the ONR through MURI award N000142112431. The work of L. F. O. Chamon is supported by the Agence Nationale de la Recherche (ANR) project ANR-25-CE23-3477-01, by a Hi!PARIS chair, funded in part by the France 2030 program (ANR-23-IACL-0005), and by the Shiing-Shen Chern Young Faculty Award from the Fondation de l'École polytechnique.

Appendix A. Notation

In this section we collect the notation used throughout the paper.

A.1 Abbreviations

1. w.h.p. stands for "with high probability"
2. i.i.d. stands for "independent identically distributed"
3. PSD stands for "positive semi-definite"
4. MNI stands for "minimum norm interpolating solution".

A.2 Basic notation

1. Symbol $:=$ denotes a definition of a quantity, for example $b := a + 1$ means that b is defined as $a + 1$.
2. $a \wedge b := \min(a, b)$.
3. $a \vee b := \max(a, b)$.
4. $a_+ := a \vee 0$.
5. $\boldsymbol{e}_i$ is the i-th coordinate (basis) vector in $\mathbb{R}^p$.
6. $a \approx b$ denotes an informal statement that a and b are w.h.p. within a constant factor of each other. Analogously, we use $a \gtrsim b$ ($a \lesssim b$) to denote "w.h.p. a is at least (at most) constant times b"
7. $a \gg b$ ($a \ll b$) denotes an informal statement that "w.h.p. a is at least a large (at most a small) constant times b."
8. 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.
9. $\mathrm{diag}(\boldsymbol{v})$ denotes the diagonal matrix in $\mathbb{R}^{d\times d}$ whose diagonal elements are elements of $\boldsymbol{v} \in \mathbb{R}^d$.
10. $\mathbf{0}_d \in \mathbb{R}^d$ is the vector of all zeros.
11. $\mathbf{1}_d \in \mathbb{R}^d$ is the vector of all ones.
12. $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix.
13. For a symmetric matrix $\boldsymbol{M}$, $\mu_i(\boldsymbol{M})$ denotes its i-th eigenvalue.
14. For any PSD matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ and any $\boldsymbol{u} \in \mathbb{R}^d$, we define $\|\boldsymbol{u}\|_{\boldsymbol{M}} := \sqrt{\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u}}$.
15. For any matrix $\boldsymbol{M} \in \mathbb{R}^{n\times p}$, we denote its pseudo-inverse by $\boldsymbol{M}^\dagger \in \mathbb{R}^{p\times n}$.
16. $\mathbb{P}(\mathscr{A})$ denotes the probability of an event $\mathscr{A}$.

17. $\mathbb{E}[\xi]$ denotes expectation of a random variable ξ.

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

A.3 Problem setting

1. $\boldsymbol{y} \in \{-1, 1\}^n$ is the vector of true class labels, whose components are i.i.d. Rademacher random variables.

2. $\hat{\boldsymbol{y}} \in \{-1, 1\}^n$ is a vector of observed labels, that is, $\hat{\boldsymbol{y}}$ is obtained from $\boldsymbol{y}$ by flipping the sign of each of its coordinates independently with probability η.

3. $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean of positive cluster, $-\boldsymbol{\mu}$ is the mean of negative cluster.

4. $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is the covariance matrix for both clusters.

5. $\{\lambda_i\}_{i=1}^p$ denote the eigenvalues of $\boldsymbol{\Sigma}$ in non-increasing order. In the eigenbasis of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$.

6. $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ is a matrix with i.i.d. isotropic rows.

7. $\boldsymbol{Q} := \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{n \times p}$.

8. $\boldsymbol{X} := \boldsymbol{y}\boldsymbol{\mu}^\top + \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{n \times p}$ is the matrix whose rows are the observed covariates.

9. $\boldsymbol{\nu} := \boldsymbol{Q}\boldsymbol{\mu}$.

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

A.4 Ridge regression and minimum norm interpolating solutions

1. $\lambda \in \mathbb{R}$ is a ridge regularization parameter.

2. $\boldsymbol{w}_{\text{ridge}} := \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_n)^{-1}\hat{\boldsymbol{y}}$—ridge regression solution.

3. $\boldsymbol{w}_{\text{MNI}} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\hat{\boldsymbol{y}}$—MNI (Minimum Norm Interpolating solution).

4. $\boldsymbol{D}_{\hat{\boldsymbol{y}}} := \operatorname{diag}(\hat{\boldsymbol{y}})$.

5. $\tilde{\boldsymbol{w}}_{\text{MNI}} := \boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = \operatorname{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^n} \|\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}\boldsymbol{\alpha}\|$ s.t. $\boldsymbol{\alpha}^\top \mathbf{1}_n = 1$,—an alternative definition of MNI direction, this vector has the same direction as MNI but different norm.

6. The following expressions give the notation for MNI and its alternative definition for the particular case of $\eta = 0$.

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

A.5 Objects from geometric derivations

1. For a linear space $\mathcal{A} \subseteq \mathbb{R}^p$, we denote its orthogonal complement by $\mathcal{A}^\perp$.
2. $\mathcal{Q}$ denotes the linear span of the columns of $\boldsymbol{Q}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$.
3. $\mathcal{Q}_A$ denotes the linear space parallel to the affine span of the columns of $\boldsymbol{Q}^\top \boldsymbol{D}_{\hat{\boldsymbol{y}}}$.
4. For any $\boldsymbol{v} \in \mathbb{R}^p$, $\boldsymbol{v}_\parallel$ denotes the projection of $\boldsymbol{v}$ on $\mathcal{Q}$, and $\boldsymbol{v}_\perp$ denotes the projection of $\boldsymbol{v}$ on $\mathcal{Q}^\perp$.
5. $\xi := \frac{\boldsymbol{y}^\top \boldsymbol{A}^{-1} \hat{\boldsymbol{y}}}{\boldsymbol{y}^\top \boldsymbol{A}^{-1} \boldsymbol{y}}$.
6. $\tilde{\boldsymbol{y}} := \hat{\boldsymbol{y}} - \xi \boldsymbol{y}$.

A.6 Separation of the first k components

1. For any $k \in \{0, 1, \dots, p\}$ and any matrix $\boldsymbol{M} \in \mathbb{R}^{n \times p}$, we denote by $\boldsymbol{M}_{0:k}$ the matrix comprised of the first k columns of $\boldsymbol{M}$.[5]
2. $\boldsymbol{M}_{k:\infty}$ denotes the matrix comprised of the last $p-k$ columns of $\boldsymbol{M}$.
3. $\boldsymbol{u}_{0:k}$ denotes the vector comprised of the first k components of $\boldsymbol{u} \in \mathbb{R}^p$ and by $\boldsymbol{u}_{k:\infty}$ denotes the remaining components.
4. $\boldsymbol{\Sigma}_{0:k} := \operatorname{diag}(\lambda_1, \dots, \lambda_k)$ and $\boldsymbol{\Sigma}_{k:\infty} := \operatorname{diag}(\lambda_{k+1}, \dots, \lambda_p)$.
5. $\boldsymbol{A}_k := \boldsymbol{Q}_{k:\infty} \boldsymbol{Q}_{k:\infty}^\top + \lambda \boldsymbol{I}_n$.

A.7 High probability events

1. For any $k \in \{0, 1, \dots, p-1\}$ and $L \geq 1$, we $\mathscr{A}_k(L)$ denotes the following event

$$\mathscr{A}_k(L) := \left\{ \frac{1}{L}\Big(\lambda + \sum_{i>k} \lambda_i\Big) \leq \mu_n(\boldsymbol{A}_k) \leq \mu_1(\boldsymbol{A}_k) \leq L\Big(\lambda + \sum_{i>k} \lambda_i\Big) \right\}.$$

2. For any $k \in \{0, 1, \dots, p-1\}$ and $c_B > 0$, we let $\mathscr{B}_k(c_B)$ be the event on which all of the following hold:

$$\begin{aligned}
&\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \leq c_B n \quad \text{and} \quad \mu_n(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \geq n/c_B, \\
&\|\boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\|^2 \leq c_B n \|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}, \\
&\operatorname{tr}(\boldsymbol{Q}_{k:\infty} \boldsymbol{\Sigma}_{k:\infty} \boldsymbol{Q}_{k:\infty}^\top) \leq c_B n \sum_{i>k} \lambda_i^2, \\
&\operatorname{tr}(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \leq c_B n k, \\
&\|\boldsymbol{Q}_{k:\infty} \boldsymbol{\Sigma}_{k:\infty} \boldsymbol{Q}_{k:\infty}^\top\| \leq c_B \left(\sum_{i>k} \lambda_i^2 + n \lambda_{k+1}^2 \right).
\end{aligned}$$

5. When $k = 0$, this matrix is empty and all the terms involving the indices $0 : k$ become zero.

A.8 Quantities from the main bounds

1. $\sigma_\eta := 1/\sqrt{\ln \frac{3+\eta^{-1}}{2}}$.
2. $\Lambda := \lambda + \sum_{i>k} \lambda_i$.
3. $V := n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{1:k}^{-1} + \mathbf{I}_k\right)^{-2}\right) + \Lambda^{-2} n \sum_{i>k} \lambda_i^2$.
4. $\Delta V := \frac{1}{n} \wedge \frac{n\lambda_{k+1}^2}{\Lambda^2} + \frac{n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2}{\Lambda^2}$.
5. $B := n^{-2}\Lambda^2 \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{1:k}^{-1} + \mathbf{I}_k\right)^{-1} \boldsymbol{\Sigma}_{1:k}^{-1/2} \boldsymbol{\mu}_{1:k} \right\|^2 + \|\boldsymbol{\mu}_{k+1:\infty}\|^2_{\boldsymbol{\Sigma}_{k+1:\infty}}$.
6. $\Diamond^2 := n\Lambda^{-2} B$.
7. $M := \frac{\Lambda}{n} \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{1:k}^{-1} + \mathbf{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{1:k}^{-1/2} \boldsymbol{\mu}_{1:k} \right\|^2 + \|\boldsymbol{\mu}_{k+1:\infty}\|^2$.
8. $N := n\Lambda^{-1} M$.
9. $\alpha_\varepsilon := \inf\left\{\alpha \in \mathbb{R} : \mathbb{P}\left(\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{\mathrm{ridge}}}{\|\boldsymbol{w}_{\mathrm{ridge}}\|_{\boldsymbol{\Sigma}}} < \alpha\right) > \varepsilon\right\}$.

A.9 Alternative forms of main quantities

1. $k^* := \min\left\{\kappa \in \{0, 1, \dots, k\} : \lambda + \sum_{i>\kappa} \lambda_i \geq n\lambda_{\kappa+1}\right\}$.
2. $\Lambda_* := \lambda + \sum_{i>k^*} \lambda_i$.
3. $V_* := \frac{k^*}{n} + \Lambda_*^{-2} n \sum_{i>k^*} \lambda_i^2$.
4. $\Diamond_*^2 := n^{-1} \|\boldsymbol{\mu}_{1:k^*}\|^2_{\boldsymbol{\Sigma}_{1:k^*}^{-1}} + n\Lambda_*^{-2} \|\boldsymbol{\mu}_{k^*+1:\infty}\|^2_{\boldsymbol{\Sigma}_{k+1:\infty}}$.
5. $N_* := \|\boldsymbol{\mu}_{1:k^*}\|^2_{\boldsymbol{\Sigma}_{1:k^*}^{-1}} + n\Lambda_*^{-1} \|\boldsymbol{\mu}_{k^*+1:\infty}\|^2$.
6. $N_a := \sum_i \frac{\mu_i^2}{\lambda_i + \Lambda/n}$.
7. $V_a := \sum_i \frac{\lambda_i^2/n}{(\lambda_i + \Lambda/n)^2}$.
8. ${\Diamond_a}^2 := \sum_i \frac{\lambda_i \mu_i^2/n}{(\lambda_i + \Lambda/n)^2}$.

Appendix B. Proof outline and different forms of the main quantities

B.1 Lower bound proof sketch

On a high level we follow the same logic as Tsigler and Bartlett (2023): start by using algebraic formulas for the ridge/MNI solutions to disentangle the contribution of the $0:k$ from the $k:\infty$ components, then estimate each quantity using concentration inequalities. This is, however, more challenging for our mixture model since the clusters are not zero-mean (they are centered at $\pm\boldsymbol{\mu}$). Algebraically, the rank-one modification in $\boldsymbol{X} = \boldsymbol{Q} + \boldsymbol{y}\boldsymbol{\mu}^\top$ hinders the straightforward application of the machinery from (Tsigler and Bartlett, 2023).

Hence, 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 involves 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 Lemma 35 (Appendix C). We then proceed to deriving sharp bounds for each of its terms. Note that there are two independent sources of randomness in our setting: the matrix $\boldsymbol{Q}$ and the labels $(\boldsymbol{y}, \hat{\boldsymbol{y}})$. We start by addressing the second source, making high probability statements over the draw of $(\boldsymbol{y}, \hat{\boldsymbol{y}})$ conditionally on $\boldsymbol{Q}$ in Lemma 49 (Appendix F). As a result, the problem now reduces to bounding expressions that only depend on $\boldsymbol{Q}$.

At this point we can apply the ideas developed in (Tsigler and Bartlett, 2023) with some modifications. As in that paper, we start by deriving algebraic bounds that hold almost surely on the event that the matrix $\boldsymbol{A}_k$ is PD (Lemma 50, Appendix G). Some of the terms that we bound there have already appeared in (Tsigler and Bartlett, 2023), namely $\|\boldsymbol{\mu}_{\gtrsim}\|^2_{\boldsymbol{\Sigma}}$ and $\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})$ are respectively the bias and variance from (Tsigler and Bartlett, 2023). Though we could, in principle, reuse the results in that paper, we use a slightly different derivation to obtain them in a new form.

Explicitly, we directly use the Sherman-Morrison-Woodbury identity to obtain our algebraic decompositions (Lemma 51, Appendix G), the most important embodiment of which is $\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k} = \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}$. We then use the results in Lemma 52 to replace $\boldsymbol{\Sigma}_{0:k}^{1/2}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{1/2}$ 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. This is in contrast to the strategy used in (Tsigler and Bartlett, 2023), which uses the fact 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 $k = k^*$. Our technique yields sharp results for any choice of k and upper bounds that have the same form as lower bounds. In Tsigler and Bartlett (2023), upper and lower bounds have different forms and a separate conversion is needed to show that they coincide for $k = k^*$. We pay for that, however, with a bulkier form of the bounds.

The bounds from Lemma 50 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 in the bounds is done in Lemma 53 (Appendix H). We conclude the proof of Theorem 8 in Appendix I by first combining the bounds from Lemmas 49 and 53 in Lemma 54 and then plugging the result into the algebraic formulas from Lemma 35.

B.2 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. The 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 $\lozenge$, 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 the distribution 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 $|a + \varepsilon b|$.

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

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

where $\bar{\boldsymbol{\nu}} = \bar{\boldsymbol{Q}}\boldsymbol{\mu}$, $\bar{\boldsymbol{\mu}}_{\perp} = (\boldsymbol{I}_p - \bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{Q}})\boldsymbol{\mu}$, and $\bar{S}$ is a scalar, which is non-negative with high probability. The proof of Theorem 8 already provides a sharp high-probability bound on the term $\bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}} \boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\perp}$, 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 $\lozenge$. 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 55. We then use them in Lemma 56 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

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

The term $|\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}}| \|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}$ is dominated by others, so we can reuse an upper bound on it from the proof of Theorem 8. When it comes to the remaining terms, note that $\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}} = \varepsilon_y \bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1} \boldsymbol{y}$, while $\bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}} \bar{\boldsymbol{\mu}}_{\perp} = \boldsymbol{y}^\top \boldsymbol{A}^{-1} \boldsymbol{y} \bar{\boldsymbol{\mu}}_{\perp}$, 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 $\bar{\boldsymbol{\mu}}_{\lesssim}$. 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 57.

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 58 we make high-probability statements with respect to the draw of $\boldsymbol{y}$, and in Lemma 59 we make high probability statements with respect to the draw of $\boldsymbol{Q}$. We put together the lower bound on the denominator in Lemma 60 and combine it with the upper bound on the numerator in Theorem 9, whose proof is given in Appendix J.3.

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

Appendix C. Formulas for the solution

In this section we derive the explicit formula for $\boldsymbol{w}_{\text{ridge}}$, which operates with the 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 Cao et al. (2021), who, in turn, took it from Wang and Thrampoulidis (2021).

Lemma 34 (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

$$\begin{aligned}
S\boldsymbol{w}_{MNI} =& \left[(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \|\boldsymbol{\mu}_\perp\|^2 \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right] \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \\
&+ \left[(1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})(1 - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}) - \|\boldsymbol{\mu}_\perp\|^2 \Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right] \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
&+ \left[\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\right] \boldsymbol{\mu}_\perp, \\
S\boldsymbol{\mu}^\top \boldsymbol{w}_{MNI} =& \boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\|\boldsymbol{\mu}_\perp\|^2 + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \\
=& \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
&+ \boldsymbol{y}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\|\boldsymbol{\mu}_\perp\|^2 + (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}.
\end{aligned}$$

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 a 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 $\boldsymbol{w}_{\text{MNI}}$ 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 $\boldsymbol{P}_{\boldsymbol{Q}} := \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}$. For any $\boldsymbol{v} \in \mathbb{R}^p$ denote the projection of $\boldsymbol{v}$ on $\mathcal{Q}$ as $\boldsymbol{v}_\parallel$ and the projection of $\boldsymbol{v}$ on $\mathcal{Q}^\perp$ as $\boldsymbol{v}_\perp$:

$$\boldsymbol{v}_\perp := (\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{Q}})\boldsymbol{v}, \quad \boldsymbol{v}_\parallel := \boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v} = \boldsymbol{v} - \boldsymbol{v}_\perp.$$

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

$$\begin{aligned}\boldsymbol{X}\boldsymbol{w} =&(\boldsymbol{Q} + \boldsymbol{y}\boldsymbol{\mu}_\parallel^\top)\boldsymbol{w}_\parallel + \boldsymbol{y}\boldsymbol{\mu}_\perp^\top\boldsymbol{w}_\perp \\ =&\boldsymbol{Q}\boldsymbol{w}_\parallel + (\boldsymbol{\mu}_\parallel^\top\boldsymbol{w}_\parallel + \alpha\|\boldsymbol{\mu}_\perp\|^2)\boldsymbol{y}\end{aligned}$$

Recall that we want to find the minimum norm interpolator for labels $\hat{\boldsymbol{y}}$, that is such $\hat{\boldsymbol{w}}$ that

$$\boldsymbol{Q}\hat{\boldsymbol{w}}_\parallel + (\boldsymbol{\mu}_\parallel^\top\hat{\boldsymbol{w}}_\parallel + \alpha\|\boldsymbol{\mu}_\perp\|^2)\boldsymbol{y} = \hat{\boldsymbol{y}}.$$

Denote $\beta := \boldsymbol{\mu}_\parallel^\top\hat{\boldsymbol{w}}_\parallel + \alpha\|\boldsymbol{\mu}_\perp\|^2$. We see that $\hat{\boldsymbol{w}}_\parallel$ is such vector in $\mathcal{Q}$ that $\boldsymbol{Q}\hat{\boldsymbol{w}}_\parallel = \hat{\boldsymbol{y}} - \beta\boldsymbol{y}$. 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

$$\hat{\boldsymbol{w}}_\parallel = \boldsymbol{Q}^\top(\underbrace{\boldsymbol{Q}\boldsymbol{Q}^\top}_{\boldsymbol{A}})^{-1}(\hat{\boldsymbol{y}} - \beta\boldsymbol{y}).$$

Thus far, we learned that for some scalars α, β it holds that

$$\begin{aligned}\hat{\boldsymbol{w}}_\perp =&\alpha\boldsymbol{\mu}_\perp, \\ \hat{\boldsymbol{w}}_\parallel =&\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} - \beta\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}, \\ \beta =&\boldsymbol{\mu}_\parallel^\top\hat{\boldsymbol{w}}_\parallel + \alpha\|\boldsymbol{\mu}_\perp\|^2 \\ =&\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} - \beta\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y} + \alpha\|\boldsymbol{\mu}_\perp\|^2.\end{aligned}$$

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 at 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 $\|\hat{\boldsymbol{w}}\|^2 = \|\hat{\boldsymbol{w}}_\parallel\|^2 + \|\hat{\boldsymbol{w}}_\perp\|^2$ 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 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

$$\hat{\boldsymbol{w}}_\parallel = \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} + (1-\beta)\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}, \quad \hat{\boldsymbol{w}}_\perp = \alpha\boldsymbol{\mu}_\perp.$$

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 \Big[(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}\Big] \\
&+ \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 35 (Explicit formulas for the ridge solution) *Denote*

$$\begin{aligned}
\Delta\boldsymbol{y} :=& \hat{\boldsymbol{y}} - \boldsymbol{y}, \\
\boldsymbol{\mu}_{\perp\!\!\!\approx} :=& (\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu}, \\
S :=& (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp\!\!\!\approx} \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}.
\end{aligned}$$

Then for any λ such that the matrix $\boldsymbol{A}$ is PD the following holds:

$$\begin{aligned}
S\boldsymbol{w}_{ridge} =& \left[(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp\!\!\!\approx}\right] \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \\
&+ \left[(1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})(1-\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}) - \Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp\!\!\!\approx}\right] \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
&+ \left[\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\right] \boldsymbol{\mu}_{\perp\!\!\!\approx}, \\
S\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} =& \boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp\!\!\!\approx} + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \\
=& \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp\!\!\!\approx} + (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \\
&+ \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}) + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y} \cdot \boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp\!\!\!\approx}.
\end{aligned}$$

In particular, when $\hat{\boldsymbol{y}} = \boldsymbol{y}$

$$S\boldsymbol{w}_{ridge} = (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}_{\perp\!\!\!\approx}, \tag{41}$$

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} = (1+\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp\!\!\!\approx}. \tag{42}$$

Proof First of all, we obtain formulas for the particular case when $\lambda = 0$ in Lemma 34 below. These formulas can be extended to the case of positive λ by a standard trick. Recall the definitions:

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

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 34 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:

$$\begin{aligned}
\check{\boldsymbol{Q}} :=& [\boldsymbol{Q}, \sqrt{\lambda}\boldsymbol{I}_n], \\
\check{\boldsymbol{A}} :=& \check{\boldsymbol{Q}}\check{\boldsymbol{Q}}^\top = \boldsymbol{Q}\boldsymbol{Q}^\top + \lambda \boldsymbol{I}_n = \boldsymbol{A}, \\
\check{\boldsymbol{\nu}} :=& \check{\boldsymbol{Q}}\check{\boldsymbol{\mu}} = \boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{\nu}, \\
\check{\boldsymbol{\mu}}_\perp :=& (\boldsymbol{I}_{p+n} - \check{\boldsymbol{Q}}^\top \check{\boldsymbol{A}}^{-1}\check{\boldsymbol{Q}})\check{\boldsymbol{\mu}} = \begin{pmatrix} (\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu} \\ -\sqrt{\lambda}\boldsymbol{A}^{-1}\boldsymbol{\nu} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\perp} \\ -\sqrt{\lambda}\boldsymbol{A}^{-1}\boldsymbol{\nu} \end{pmatrix}
\end{aligned}$$

The only thing that is not straightforward to plug in is $\|\check{\boldsymbol{\mu}}_\perp\|^2$, which we derive next:

$$\begin{aligned}
\|\check{\boldsymbol{\mu}}_\perp\|^2 =& \|(\boldsymbol{I}_p - \boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu}\|^2 + \lambda\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|^2 \\
=& \|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{\nu} + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{\nu} + \lambda \boldsymbol{\nu}^\top \boldsymbol{A}^{-1} p \boldsymbol{A}^{-1}\boldsymbol{\nu} \\
=& \|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{\nu} + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}(\underbrace{\boldsymbol{Q}\boldsymbol{Q}^\top + \lambda \boldsymbol{I}_n}_{\boldsymbol{A}})\boldsymbol{A}^{-1}\boldsymbol{\nu} \\
=& \|\boldsymbol{\mu}\|^2 - \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{\nu} \\
=& \boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp}.
\end{aligned}$$

Now we can obtain the result for $\lambda \geq 0$: Plugging all those objects in Lemma 34 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 (41) 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 $\{\lambda \in \mathbb{R} : \lambda > -\mu_n(-\mu_n(\boldsymbol{Q}\boldsymbol{Q}^\top))\}$. ■

Appendix D. General probabilistic results

First of all, we define sub-Gaussian norms of random variables and vectors.

Definition 36 *For any centered random variable v we define its sub-Gaussian norm as*

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

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.

Lemma 37 *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 36, 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

$$\mathbb{E}\exp\left((\xi^2-\eta)^2/t^2\right) = \eta e^{(1-\eta)^2/t^2} + (1-\eta)e^{\eta^2/t^2} \le 2.$$

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 $\eta^{-1} \geq 1$, we have

$$\eta^{-2} \ln \frac{3}{2} = \ln(3e^{\eta^{-2}}/2) \geq \ln(3(1+\eta^{-2})/2) \geq \ln(3(1+\eta^{-1})/2),$$

so $\|\xi^2 - \eta\|_{\psi_2} \leq t = 1/\sqrt{\ln \frac{3+\eta^{-1}}{2}}$.

Finally, we compare two bounds that we obtained:

$$1 + \eta^{-1} - \frac{3+\eta^{-1}}{2} = \frac{\eta^{-1}-1}{2} \geq 0.$$

We see that $1/\sqrt{\ln(1+1/\eta)} \leq 1/\sqrt{\ln \frac{3+\eta^{-1}}{2}}$. ■

Lemma 38 *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 $1 - 2e^{-t^2/c}$*

$$\sqrt{\sum_i \eta_i^2} \leq \sigma(\sqrt{n} + t)$$

Proof We basically repeat the proof of Theorem 3.1.1 from Vershynin (2018), 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 $v \leq \sigma \leq 1$.

As in the proof of Theorem 3.1.1 from Vershynin (2018), 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 Vershynin (2018), we get that for some absolute constant $c > 0$ and any $u \geq 0$ with probability at least $1 - 2\exp(-cn(u \wedge u^2))$

$$n^{-1} \sum_i \eta_i^2 \leq v^2 + u \leq 1 + u \leq (1 + (\sqrt{u} \wedge u))^2.$$

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 $1 - 2e^{-ct^2}$

$$\sqrt{n^{-1} \sum_i \eta_i^2} \leq 1 + t/\sqrt{n}.$$

■

Lemma 39 (Weakened 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 $s > 0$*

$$\mathbb{P}\left\{ |\boldsymbol{\varepsilon}^\top \boldsymbol{M} \boldsymbol{\varepsilon} - v^2 tr(\boldsymbol{M})| > \sigma^2 \max(\sqrt{s}\|\boldsymbol{M}\|_F, s\|\boldsymbol{M}\|) \right\} \leq 2\exp\{-s/c\}.$$

Proof This is basically a rewriting of Theorem 6.2.1 (Hanson-Wright inequality) in Vershynin (2018). According to that theorem, for some absolute constant c for any $t > 0$,

$$\mathbb{P}\left\{|\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} - \mathbb{E}_{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon}| \geq t\right\} \leq 2\exp\left(-c^{-1}\min\left\{\frac{t^2}{\|\boldsymbol{M}\|_F^2\sigma^4}, \frac{t}{\|\boldsymbol{M}\|\sigma^2}\right\}\right),$$

where $\mathbb{E}_\varepsilon$ 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 40 (Weakened Hanson-Wright for PSD matrices) *In the setting of Lemma 39 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 39 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 39. 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. ■

Lemma 41 (Lemma 24 from Tsigler and Bartlett (2023)) *Suppose that $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{n\times p}$ is a matrix with i.i.d. isotropic sub-Gaussian rows with sub-Gaussian constant σ. 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 42 *There exists a constant c that only depends on σ_x such that with probability at least $ce^{-n/c}$*

$$\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\| \le c\left(\sum_{i>k}\lambda_i^2 + n\lambda_{k+1}^2\right).$$

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 41 for $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}_{k:\infty}$ and $\boldsymbol{M} = \boldsymbol{\Sigma}_{k:\infty}^2$. ■

Lemma 43 *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}*

$$|\boldsymbol{v}^\top\boldsymbol{y}| \ge c^{-1}\|\boldsymbol{v}\|.$$

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 $t > 0$

$$\mathbb{P}(\xi > t) \le 2e^{-t^2/c_2}.$$

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 $c_4 > c_3$

$$\begin{aligned}
1 =& \mathbb{E}[\xi^2] \\
=& \int_0^\infty \mathbb{P}(|\xi|^2 > t)\,dt \\
\le& \int_0^{c_3} 1\,dt + \int_{c_3}^{c_4}\mathbb{P}(|\xi|^2 > t)\,dt + 2\int_{c_4}^\infty e^{-t/c_2}\,dt \\
\le& c_3 + (c_4 - c_3)\mathbb{P}(\xi^2 > c_3) + 2c_2e^{-c_4/c_2}.
\end{aligned}$$

We see that

$$\mathbb{P}(\xi^2 > c_3) \ge \frac{1 - c_3 - 2c_2e^{-c_4/c_2}}{c_4}$$

Taking c_4 to be a large enough absolute constant, and c_3—small enough, yields the result. ■

Lemma 44 *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}*

$$|\boldsymbol{y}^\top\boldsymbol{M}\boldsymbol{y}| \ge c^{-1}tr(\boldsymbol{M}).$$

Proof First of all, since $\boldsymbol{y}$ has i.i.d. centered coordinates with sub-Gaussian norms bounded by an absolute constant, by Corollary 40 for some absolute constant c_1 and any $s > 0$

$$\mathbb{P}(\boldsymbol{y}^\top \boldsymbol{M}\boldsymbol{y} > c_1(\operatorname{tr}(\boldsymbol{M}) + s\|\boldsymbol{M}\|)) \le 2e^{-s/c_1}.$$

Denote $\xi = \boldsymbol{y}^\top \boldsymbol{M}\boldsymbol{y}/\operatorname{tr}(\boldsymbol{M})$. 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 43. We write for some small c_2, and large $c_3 > c_1$

$$\begin{aligned}
1 =& \mathbb{E}[\xi] \\
=& \int_0^\infty \mathbb{P}(|\xi| > t)\, dt \\
\le& \int_0^{c_2} 1\, dt + \int_{c_2}^{c_3} \mathbb{P}(|\xi| > t)\, dt + 2\int_{c_3}^{\infty} e^{-(t/c_1-1)/c_1}\, dt \\
\le& c_2 + (c_3 - c_2)\mathbb{P}(\xi > c_2) + 2c_1 e^{-(c_3/c_1 - 1)/c_1}.
\end{aligned}$$

$$\mathbb{P}(\xi > c_2) \ge \frac{1 - c_2 - 2c_1 e^{-(c_3/c_1-1)/c_1}}{c_3 - c_2}.$$

Taking c_2 to be a small enough constant, and c_3—large enough, yields the result. ∎

Appendix E. Some important relations

Lemma 45 (Relations between the main quantities) *Suppose that*

$$k \le n \quad \text{and} \quad \Lambda > n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}. \tag{14}$$

Then, $n\lozenge^2 \le N$, $n\lozenge^2 \le N\sqrt{n\Delta V}$, $V \le 2$, $\Delta V \le \dfrac{3}{n}$, *and* $\Delta V \le 4V$.

Proof

For the first inequality, we write

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

where we used that $\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\right\| \leq 1$ in the second line, and $\Lambda > n\lambda_{k+1}$ in the third line. Alternatively, we could use $\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\right\| \leq n\Lambda^{-1}\lambda_1$ in the second line to obtain

$$\begin{aligned}\lozenge^2 \leq& n\Lambda^{-2}\lambda_{k+1}\|\boldsymbol{\mu}_{k:\infty}\|^2 + n\Lambda^{-1}\lambda_1 \cdot n^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\\ \leq& n\Lambda^{-2}\lambda_1\left(\|\boldsymbol{\mu}_{k:\infty}\|^2 + n^{-1}\Lambda\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2\right)\\ =& n\Lambda^{-2}\lambda_1 M,\end{aligned}$$

which means that

$$\lozenge^2 \leq \Lambda^{-1}M \wedge n\Lambda^{-2}\lambda_1 M = \left(\frac{1}{\sqrt{n}} \wedge \frac{\sqrt{n}\lambda_1}{\Lambda}\right)\sqrt{n}\Lambda^{-1}M \leq \Lambda^{-1}M\sqrt{n\Delta V}.$$

Now let's upper bound V:

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

where we used Equation (14) 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

$$n\Delta V \leq 1 + \frac{n^2\lambda_{k+1}^2 + n\sum_{i>k}\lambda_i^2}{\Lambda^2} \leq 3,$$

where the last transition follows directly from Equation (14).

Finally, let's compare V and ΔV. In case $k=0$ we get

$$4V = 4\Lambda^{-2}n\sum_i \lambda_i^2 \geq \Lambda^{-2}\left(2n\lambda_1^2 + \sum_i\lambda_i^2\right) = \Delta V.$$

If $k>0$, we have

$$\begin{aligned}4V =& 4n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right) + 4\Lambda^{-2}n\sum_{i>k}\lambda_i^2\\ \geq& 4n^{-1}\frac{1}{(1+\Lambda n^{-1}\lambda_1^{-1})^2} + \Lambda^{-2}n\lambda_{k+1}^2 + \Lambda^{-2}\sum_{i>k}\lambda_i^2\\ \geq& n^{-1}\frac{1}{(1\vee\Lambda n^{-1}\lambda_1^{-1})^2} + \Lambda^{-2}n\lambda_{k+1}^2 + \Lambda^{-2}\sum_{i>k}\lambda_i^2\\ =& \Delta V.\end{aligned}$$

■

Lemma 46 (Bounds via k^*) *Suppose that*

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

Define

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

Then

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

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 $n\Lambda_*^{-1}M_*$.

$$\begin{aligned}
n\Lambda^{-1}M &= \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k} \right\|^2 + n\Lambda^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2 \\
&= \sum_{i=1}^{k} \frac{\mu_i^2}{\lambda_i(1+\lambda_i^{-1}n^{-1}\Lambda)} + n\Lambda^{-1}\sum_{i>k}\mu_i^2 \\
&= \sum_{i=1}^{k^*} \frac{\mu_i^2}{\lambda_i + n^{-1}\Lambda} + \sum_{i=k^*+1}^{k} \frac{\mu_i^2}{\lambda_i + n^{-1}\Lambda} + \sum_{i>k}\frac{\mu_i^2}{n^{-1}\Lambda} \\
&\begin{cases} \geq \sum_{i=1}^{k^*} \frac{\mu_i^2}{2\lambda_i} + \sum_{i=k^*+1}^{k} \frac{\mu_i^2}{2n^{-1}\Lambda_*} + \sum_{i>k}\frac{\mu_i^2}{n^{-1}\Lambda_*}, \\ \leq \sum_{i=1}^{k^*} \frac{\mu_i^2}{\lambda_i} + \sum_{i=k^*+1}^{k} \frac{\mu_i^2}{n^{-1}\Lambda_*/2} + \sum_{i>k}\frac{\mu_i^2}{n^{-1}\Lambda_*/2}, \end{cases}
\end{aligned}$$

where we plugged in $n^{-1}\Lambda \leq \lambda_i$ 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

$$n\Lambda_*^{-1}M_* = \sum_{i=1}^{k^*}\frac{\mu_i^2}{\lambda_i} + \sum_{i>k^*}\frac{\mu_i^2}{n^{-1}\Lambda_*},$$

the above implies that $2n\Lambda_*^{-1}M_* \geq n\Lambda^{-1}M \geq n\Lambda_*^{-1}M_*/2$.

2. $\lozenge$ vs $\lozenge_*$.

$$\begin{aligned}
n\lozenge^2 &= \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k} \right\|^2 + n^2\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
&= \sum_{i=1}^{k} \frac{\mu_i^2}{\lambda_i(1+\lambda_i^{-1}n^{-1}\Lambda)^2} + n^2\Lambda^{-2}\sum_{i>k}\lambda_i\mu_i^2 \\
&= \sum_{i=1}^{k^*} \frac{\lambda_i\mu_i^2}{(\lambda_i + n^{-1}\Lambda)^2} + \sum_{i=k^*+1}^{k} \frac{\lambda_i\mu_i^2}{(\lambda_i + n^{-1}\Lambda)^2} + \sum_{i>k}\frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda)^2} \\
&\begin{cases} \geq \sum_{i=1}^{k^*} \frac{\lambda_i\mu_i^2}{(2\lambda_i)^2} + \sum_{i=k^*+1}^{k} \frac{\lambda_i\mu_i^2}{(2n^{-1}\Lambda_*)^2} + \sum_{i>k}\frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda_*)^2}, \\ \leq \sum_{i=1}^{k^*} \frac{\lambda_i\mu_i^2}{\lambda_i^2} + \sum_{i=k^*+1}^{k} \frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda_*/2)^2} + \sum_{i>k}\frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda_*/2)^2}, \end{cases}
\end{aligned}$$

where we plugged in $n^{-1}\Lambda \leq \lambda_i$ 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

$$n\lozenge_*^2 = \sum_{i=1}^{k^*}\frac{\lambda_i\mu_i^2}{\lambda_i^2} + \sum_{i>k^*}\frac{\lambda_i\mu_i^2}{(n^{-1}\Lambda_*)^2},$$

the above implies that $4n\lozenge_*^2 \geq n\lozenge^2 \geq n\lozenge_*^2/4$.

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 47 (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.

$$\begin{aligned}
n\Lambda^{-1}M =& \left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2+n\Lambda^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2\\
=&\sum_{i=1}^{k}\frac{\mu_i^2}{\lambda_i(1+\lambda_i^{-1}n^{-1}\Lambda)}+n\Lambda^{-1}\sum_{i>k}\mu_i^2\\
=&\sum_{i=1}^{k}\frac{\mu_i^2}{\lambda_i+\Lambda/n}+\sum_{i>k}\frac{\mu_i^2}{\Lambda/n}\\
&\begin{cases}\geq\sum_i\frac{\mu_i^2}{\lambda_i+\Lambda/n}\\ \leq 2\sum_i\frac{\mu_i^2}{\lambda_i+\Lambda/n}.\end{cases}
\end{aligned}$$

$$
\begin{aligned}
n\diamondsuit^2 =& \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k} \right\|^2 + n^2\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \\
=& \sum_{i=1}^{k} \frac{\mu_i^2}{\lambda_i(1+\lambda_i^{-1}n^{-1}\Lambda)^2} + n^2\Lambda^{-2}\sum_{i>k}\lambda_i\mu_i^2 \\
=& \sum_{i=1}^{k} \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2} + \sum_{i>k}\frac{\lambda_i\mu_i^2}{(\Lambda/n)^2} \\
& \begin{cases} \geq \sum_i \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2} \\ \leq 4\sum_i \frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2}, \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
V =& n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2}n\sum_{i>k}\lambda_i^2 \\
=& n^{-1}\sum_{i=1}^{k}\frac{1}{(1+\lambda_i^{-1}n^{-1}\Lambda)^2} + \Lambda^{-2}n\sum_{i>k}\lambda_i^2 \\
=& \sum_{i=1}^{k}\frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2} + \sum_{i>k}\frac{\lambda_i^2/n}{(\Lambda/n)^2} \\
& \begin{cases} \geq \sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2} \\ \leq 4\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}, \end{cases}
\end{aligned}
$$

■

Proposition 48 (Transition order, noisy case) *Assume the conditions of Lemma 6. Fix the direction of $\boldsymbol{\mu}$ and treat its magnitude as a parameter. For any scalar quantities a, b define by $\nearrow_a^b$ the smallest magnitude of $\boldsymbol{\mu}$ for which $b \geq a$.*

For any C, for any $n > 8C^2$, one of the following two chains of inequalities holds.

- *Case 1:*

$$\nearrow_{C\diamondsuit}^{N} \leq \nearrow_{2C\sqrt{V}}^{\sqrt{n}\diamondsuit} \leq \nearrow_{2C}^{N} \leq \nearrow_{\sqrt{n}\diamondsuit}^{N\sqrt{V}} \leq \nearrow_{1}^{\diamondsuit}.$$

- *Case 2:*

$$\nearrow_{C\diamondsuit}^{N} \leq \nearrow_{\sqrt{n}\diamondsuit}^{N\sqrt{V}} \leq \nearrow_{2C}^{N} \leq \nearrow_{2C\sqrt{V}}^{\sqrt{n}\diamondsuit} \leq \nearrow_{1}^{\diamondsuit}.$$

Moreover, both cases are possible depending on $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$.

Proof

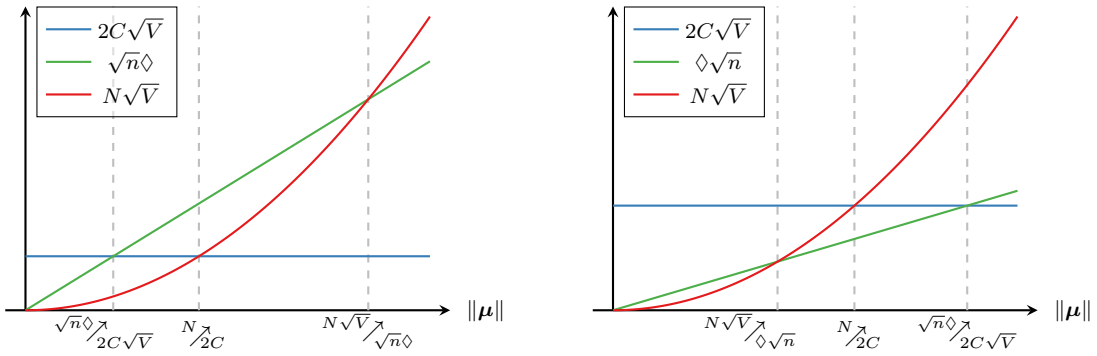
First of all, note that $2C\sqrt{V}$ does not depend on $\boldsymbol{\mu}$, $\sqrt{n}\lozenge$ has linear dependence, and $N\sqrt{V}$ has quadratic dependence. Because of that there are only two cases possible: either

$$\frac{\sqrt{n}\lozenge}{2C\sqrt{V}} \leq \frac{N}{2C} \leq \frac{N\sqrt{V}}{\sqrt{n}\lozenge},$$

or

$$\frac{N\sqrt{V}}{\sqrt{n}\lozenge} \leq \frac{N}{2C} \leq \frac{\sqrt{n}\lozenge}{2C\sqrt{V}}\,.$$

Figure 2 gives the proof by picture.



(a) Case 1: $\frac{N}{2C} > \frac{\sqrt{n}\lozenge}{2C\sqrt{V}}$ (b) Case 2: $\frac{N}{2C} < \frac{\sqrt{n}\lozenge}{2C\sqrt{V}}$

Figure 2: Two possible relations between $2C\sqrt{V}$, $\sqrt{n}\lozenge$, and $N\sqrt{V}$ depending on $\|\boldsymbol{\mu}\|$.

Now we just need to prove that the following always happen:

$$\begin{aligned}
\frac{N}{C\lozenge} &\leq \frac{\sqrt{n}\lozenge}{2C\sqrt{V}},\\
\frac{N}{C\lozenge} &\leq \frac{N\sqrt{V}}{\sqrt{n}\lozenge},\\
\frac{N\sqrt{V}}{\sqrt{n}\lozenge} &\leq \frac{\lozenge}{1},\\
\frac{\sqrt{n}\lozenge}{2C\sqrt{V}} &\leq \frac{\lozenge}{1}\,.
\end{aligned}$$

Let's do that one-by-one.

1. $\frac{N}{C\lozenge} \leq \frac{\sqrt{n}\lozenge}{2C\sqrt{V}}$.

   According to Lemma 6, $\sqrt{n}\lozenge^2 \leq N\sqrt{\Delta V} \leq 2N\sqrt{V}$, which implies that

   $$\frac{\sqrt{V}}{\lozenge\sqrt{n}} \geq \frac{\lozenge}{2N}.$$

   That is, if $\sqrt{n}\lozenge \geq 2C\sqrt{V}$ then $N \geq C\lozenge$.

2. $\frac{N}{C\lozenge} \leq \frac{N\sqrt{V}}{\sqrt{n}\lozenge}$.

   Since $V \leq 2$ and $n \geq 2C^2$, if $N\sqrt{V} \geq \sqrt{n}\lozenge$, then $N \geq C\lozenge$, so $\frac{N}{\lozenge} \leq \frac{N\sqrt{V}}{\sqrt{n}\lozenge}$.

3. $\frac{\sqrt{n}\Diamond}{2C\sqrt{V}} \leq \frac{\Diamond}{1}$:
$$\Diamond \geq 1 \Rightarrow \sqrt{n}\Diamond \geq 2\sqrt{2}C \geq 2C\sqrt{V}.$$

4. $\frac{N\sqrt{V}}{\sqrt{n}\Diamond} \leq \frac{\Diamond}{1}$: Using Lemma 6,
$$\Diamond \geq 1 \Rightarrow N\sqrt{V} \geq N\sqrt{\Delta V}/2 \geq \sqrt{n}\Diamond^2 \geq \sqrt{n}\Diamond.$$

To see that both cases in Figure 2 are possible, consider two extreme cases. First, suppose that $\boldsymbol{\mu} = m\boldsymbol{e}_1$—a vector supported on the first coordinate. We then have
$$\Diamond\sqrt{n} = \frac{\sqrt{\lambda_1}m}{\lambda_1 + \Lambda/n} \quad \text{and} \quad N = \frac{m^2}{\lambda_1 + \Lambda/n}.$$
If $\lambda_1 \gg \Lambda/n$ and $V \ll 1$, then $m = \sqrt{\lambda_1}$ implies $\Diamond\sqrt{n} \gg (1+N)\sqrt{V}$. Thus, $\Diamond\sqrt{n}$ dominates and we are in the setting of Figure 2a. On the other hand, consider $\boldsymbol{\mu} = m\boldsymbol{e}_p$—a vector supported on the last coordinate. Then,
$$\Diamond\sqrt{n} = n\Lambda^{-1}\sqrt{\lambda_p}m \quad \text{and} \quad N = n\Lambda^{-1}m^2.$$
If additionally $V \gg n\lambda_p/\Lambda$, which is possible since λ_p can be arbitrarily small, we can write
$$(1+N)\sqrt{V} = (1+n\Lambda^{-1}m^2)\sqrt{V} \gg (1+n\Lambda^{-1}m^2)\sqrt{n\Lambda^{-1}\lambda_p} \geq 2\sqrt{n\Lambda^{-1}m^2}\sqrt{n\Lambda^{-1}\lambda_p} = 2\Diamond\sqrt{n}.$$
That is, for such choice of covariance and $\boldsymbol{\mu}$, the term $\Diamond\sqrt{n}$ does not dominate the bound for $\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ in (16) for any choice of m, and thus, the case from Figure 2b must happen. ■

Appendix F. Randomness in labels

Lemma 49 (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 37, 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 Vershynin (2018)), 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 37 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 Vershynin (2018)) 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 38 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 39 (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 $\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \leq n\|\boldsymbol{A}^{-1}\|$ 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. $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2$: the bound is a direct consequence of Corollary 40, 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. $\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2$: the bound is a direct consequence of Corollary 40, applied to $\boldsymbol{M} = \boldsymbol{A}^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}$ and $\boldsymbol{\varepsilon} = \boldsymbol{y}$.

■

Appendix G. 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 50 (Algebraic decompositions) *For any $k < n$ all the following hold almost surely on the event that the matrix $\boldsymbol{A}_k$ is PD:*

1.

$$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\| \leq & \frac{\mu_1(\boldsymbol{A}_k)}{\mu_n(\boldsymbol{A}_k)} \frac{\sqrt{\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})}}{\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})} \left\| \left(\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \mu_n(\boldsymbol{A}_k) \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k \right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\| \\ & + \mu_n(\boldsymbol{A}_k)^{-1} \|\boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\|.\end{aligned}$$

2.

$$\begin{aligned}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\precsim} \geq & \frac{1}{2} \mu_n(\boldsymbol{A}_k) \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \left\| \left(\mu_n(\boldsymbol{A}_k) \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k \right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\ & + \|\boldsymbol{\mu}_{k:\infty}\|^2 - 9 \mu_n(\boldsymbol{A}_k)^{-1} \|\boldsymbol{Q}_{0:k}^\top \boldsymbol{\mu}_{k:\infty}\|^2.\end{aligned}$$

3.

$$\begin{aligned}\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\precsim} \leq & 3 \mu_1(\boldsymbol{A}_k) \mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \left\| \left(\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \mu_1(\boldsymbol{A}_k) \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k \right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\ & + \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2 \|\boldsymbol{A}_k^{-1/2} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\|^2.\end{aligned}$$

4.

$$\begin{aligned}& \|\boldsymbol{\mu}_{\precsim}\|_{\boldsymbol{\Sigma}}^2 \\ \leq & 2 \mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-2} \mu_1(\boldsymbol{A}_k)^2 \left\| \left(\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \mu_n(\boldsymbol{A}_k) \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k \right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\ & + 2 \frac{\mu_1(\boldsymbol{A}_k)^2 \mu_1(\boldsymbol{Z}_{0:k} \boldsymbol{Z}_{0:k}^\top)}{\mu_n(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^2 \mu_n(\boldsymbol{A}_k)^2} \|\boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\|^2 \\ & + 3 \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 + 3 \|\boldsymbol{Q}_{k:\infty} \boldsymbol{\Sigma}_{k:\infty} \boldsymbol{Q}_{k:\infty}^\top\| \mu_n(\boldsymbol{A}_k)^{-2} \|\boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\|^2 \\ & + 3 \|\boldsymbol{Q}_{k:\infty} \boldsymbol{\Sigma}_{k:\infty} \boldsymbol{Q}_{k:\infty}^\top\| \frac{\mu_1(\boldsymbol{A}_k)^2 \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^2 \mu_n(\boldsymbol{A}_k)^2} \\ & \times \left\| \left(\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \mu_n(\boldsymbol{A}_k) \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k \right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2.\end{aligned} \tag{43}$$

5.

$$\begin{aligned}tr(\boldsymbol{A}^{-1} \boldsymbol{Q} \boldsymbol{\Sigma} \boldsymbol{Q}^\top \boldsymbol{A}^{-1}) \leq & \frac{\mu_1(\boldsymbol{A}_k)^2 \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^2 \mu_n(\boldsymbol{A}_k)^2} \\ & \times tr\left(\left(\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \mu_n(\boldsymbol{A}_k) \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k \right)^{-2} \right) \\ & + \mu_n(\boldsymbol{A}_k)^{-2} tr(\boldsymbol{Q}_{k:\infty} \boldsymbol{\Sigma}_{k:\infty} \boldsymbol{Q}_{k:\infty}^\top).\end{aligned} \tag{44}$$

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 G gives the proof of Lemma 50.

G.1 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 51 *If $\boldsymbol{A}_k$ is invertible, then all the following hold:*

$$\boldsymbol{A}^{-1} = \boldsymbol{A}_k^{-1} - \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}, \tag{45}$$

$$\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k} = \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}, \tag{46}$$

$$\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k} = \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}. \tag{47}$$

Proof Equation (45) 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 (46) 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 (47) from (46):

$$\begin{aligned}
&\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k} \\
=& \boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1} \\
=& \boldsymbol{I}_k + \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1} - \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1} \\
=& \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}.
\end{aligned}$$

■

Another algebraic result that we will use is as follows:

Lemma 52 *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 $\boldsymbol{u} \in \mathbb{R}^k$*

$$\left\|(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})^{-1}\boldsymbol{u}\right\| \ge \beta^{-1}\left\|(\alpha^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{u}\right\|, \tag{48}$$

$$\left\|(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})^{-1}\boldsymbol{u}\right\| \le \alpha^{-1}\left\|(\beta^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{u}\right\|. \tag{49}$$

Moreover,

$$\beta^{-2} tr\left((\alpha^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-2}\right) \leq tr\left((\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})^{-2}\right) \leq \alpha^{-2} tr\left((\beta^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-2}\right). \tag{50}$$

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 (48). 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 (49).

Finally, Equation (50) is a direct consequence of Equations (48) and (49): take $\boldsymbol{g}$ to be an isotropic Gaussian vector in $\mathbb{R}^k$. Equations (48) and (49) give

$$\beta^{-2}\left\|(\alpha^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{g}\right\|^2 \leq \left\|(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{M})^{-1}\boldsymbol{g}\right\|^2 \leq \alpha^{-2}\left\|(\beta^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{g}\right\|^2.$$

Taking expectation over $\boldsymbol{g}$ yields Equation (50). ■

G.2 $\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|$

In this section we derive an upper bound on $\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\|$. Recall that

$$\begin{aligned}
\boldsymbol{\nu} &= \boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k} + \boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}, \\
\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\| &\leq \|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\| + \|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|.
\end{aligned}$$

We bound those two terms separately. For the first term we use Equation (46):

$$\begin{aligned}
\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k} =& \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k} \\
=& \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\boldsymbol{\Sigma}_{0:k}^{1/2}\left(\boldsymbol{I}_k + \boldsymbol{\Sigma}_{0:k}^{1/2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\boldsymbol{\Sigma}_{0:k}^{1/2}\right)^{-1}\boldsymbol{\mu}_{0:k} \\
=& \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\boldsymbol{\Sigma}_{0:k}^{1/2}\left(\boldsymbol{I}_k + \boldsymbol{\Sigma}_{0:k}^{1/2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\boldsymbol{\Sigma}_{0:k}^{1/2}\right)^{-1}\boldsymbol{\mu}_{0:k} \\
=& \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}.
\end{aligned}$$

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 52 together with the observation that

$$\|\boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-2}\boldsymbol{Z}_{0:k}\| \le \mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_n(\boldsymbol{A}_k)^{-2},$$
$$\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_1(\boldsymbol{A}_k)^{-1}\boldsymbol{I}_k \preceq \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k} \preceq \mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_n(\boldsymbol{A}_k)^{-1}\boldsymbol{I}_k$$

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,

$$\begin{aligned}
\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\| \le&\frac{\mu_1(\boldsymbol{A}_k)}{\mu_n(\boldsymbol{A}_k)}\frac{\sqrt{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| \\
&+\mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|.
\end{aligned}$$

G.3 $\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\not\approx}$

G.3.1 Bound from below

In this section we derive a lower bound on $\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\not\approx}$. First of all, we write

$$\begin{aligned}
\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\not\approx} =&\boldsymbol{\mu}^\top(\boldsymbol{I}_p - \boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{Q})\boldsymbol{\mu} \\
=&\boldsymbol{\mu}_{0:k}^\top(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k} \\
&+\boldsymbol{\mu}_{k:\infty}^\top(\boldsymbol{I}_{p-k} - \boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty})\boldsymbol{\mu}_{k:\infty} \\
&-2\boldsymbol{\mu}_{0:k}^\top\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}.
\end{aligned}$$

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 (47) 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 (46) to write

$$\begin{aligned}
&\left|\boldsymbol{\mu}_{0:k}^\top \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\right| \\
=&\left|\boldsymbol{\mu}_{0:k}^\top \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\right| \\
=&\left|\boldsymbol{\mu}_{0:k}^\top \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1/2} \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1/2} \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\right| \\
\leq& w \boldsymbol{\mu}_{0:k}^\top \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \boldsymbol{\mu}_{0:k} \\
&+ w^{-1} \left\| \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1/2} \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty} \right\|^2 \\
\leq& w \boldsymbol{\mu}_{0:k}^\top \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \boldsymbol{\mu}_{0:k} + w^{-1} \|\boldsymbol{A}_k^{-1/2} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\|^2,
\end{aligned}$$

where we introduced an arbitrary scalar $w > 0$ when we used AM-GM inequality. In the last line we also used the following fact:

$$\left\| \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1/2} \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1/2} \right\| \leq 1.$$

Indeed, the matrix $\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}$ is larger than the matrix $\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1/2} (\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1/2})^\top$ in the PSD sense.

We take $w = 0.25$. So far we obtained that

$$\boldsymbol{\mu}^\top \boldsymbol{\mu}_\perp \geq \frac{1}{2} \boldsymbol{\mu}_{0:k}^\top \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \boldsymbol{\mu}_{0:k} + \|\boldsymbol{\mu}_{k:\infty}\|^2 - 9 \|\boldsymbol{A}_k^{-1/2} \boldsymbol{Q}_{0:k}^\top \boldsymbol{\mu}_{k:\infty}\|^2,$$

where we also used

$$\begin{aligned}
\boldsymbol{\mu}_{k:\infty}^\top (\boldsymbol{I}_{p-k} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{k:\infty}) \boldsymbol{\mu}_{k:\infty} =& \|\boldsymbol{\mu}_{k:\infty}\|^2 - \boldsymbol{\mu}_{k:\infty}^\top \boldsymbol{Q}_{k:\infty}^\top \underbrace{\boldsymbol{A}^{-1}}_{\preceq \boldsymbol{A}_k^{-1}} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty} \\
\geq& \|\boldsymbol{\mu}_{k:\infty}\|^2 - \|\boldsymbol{A}_k^{-1/2} \boldsymbol{Q}_{0:k}^\top \boldsymbol{\mu}_{k:\infty}\|^2.
\end{aligned}$$

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

$$\begin{aligned}
&\boldsymbol{\mu}_{0:k}^\top \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \boldsymbol{\mu}_{0:k} \\
=&\boldsymbol{\mu}_{0:k}^\top \boldsymbol{\Sigma}_{0:k}^{-1/2} \left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k}\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \\
\geq&\boldsymbol{\mu}_{0:k}^\top \boldsymbol{\Sigma}_{0:k}^{-1/2} \left(\boldsymbol{\Sigma}_{0:k}^{-1} + \mu_n(\boldsymbol{A}_k)^{-1} \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k}) \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \\
=&\mu_n(\boldsymbol{A}_k) \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \left\| \left(\mu_n(\boldsymbol{A}_k) \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 .
\end{aligned}$$

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:

$$\|\boldsymbol{A}_k^{-1/2} \boldsymbol{Q}_{0:k}^\top \boldsymbol{\mu}_{k:\infty}\|^2 \leq \mu_n(\boldsymbol{A}_k)^{-1} \|\boldsymbol{Q}_{0:k}^\top \boldsymbol{\mu}_{k:\infty}\|^2$$

Combining everything together gives the bound.

$$\begin{aligned}
\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\approx} \geq& \frac{1}{2} \mu_n(\boldsymbol{A}_k) \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \left\| \left(\mu_n(\boldsymbol{A}_k) \mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\
&+ \|\boldsymbol{\mu}_{k:\infty}\|^2 - 9 \mu_n(\boldsymbol{A}_k)^{-1} \|\boldsymbol{Q}_{0:k}^\top \boldsymbol{\mu}_{k:\infty}\|^2
\end{aligned}$$

G.3.2 Bound from above

In this section we bound $\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\approx}$ from above. This is easier than bounding it from below. Indeed, recall the decomposition from the previous section:

$$\begin{aligned}
\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\approx} =& \boldsymbol{\mu}_{0:k}^\top (\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{0:k}) \boldsymbol{\mu}_{0:k} \\
&+ \boldsymbol{\mu}_{k:\infty}^\top (\boldsymbol{I}_{p-k} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{k:\infty}) \boldsymbol{\mu}_{k:\infty} \\
&- 2 \boldsymbol{\mu}_{0:k}^\top \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}.
\end{aligned}$$

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

$$\begin{aligned}
\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\not\perp} \leq & 3\boldsymbol{\mu}_{0:k}^\top \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \boldsymbol{\mu}_{0:k} + \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2\|\boldsymbol{A}_k^{-1/2} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\|^2 \\
= & 3\boldsymbol{\mu}_{0:k}^\top \boldsymbol{\Sigma}_{0:k}^{-1/2} \left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k}\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} + \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2\|\boldsymbol{A}_k^{-1/2} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\|^2 \\
\leq & 3\mu_1(\boldsymbol{A}_k)\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \left\| \left(\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \mu_1(\boldsymbol{A}_k) \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\
& + \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2\|\boldsymbol{A}_k^{-1/2} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\|^2.
\end{aligned}$$

G.4 $\|\boldsymbol{\mu}_{\not\perp}\|_{\boldsymbol{\Sigma}}^2$

The quantity $\|\boldsymbol{\mu}_{\not\perp}\|_{\boldsymbol{\Sigma}}^2$ is exactly the bias term from Tsigler and Bartlett (2023): indeed, plug in $\boldsymbol{\mu}$ instead of their θ^*, $\boldsymbol{Q}$ instead of their X, $\boldsymbol{\Sigma}$ instead of their Σ, and $\boldsymbol{A}$ instead of their A. We could, in principle, just borrow an algebraic bound from their Lemma 28. However, for the sake of completeness, and because we would like a bound in a slightly different form, we do our own derivation here.

As before, we start with the first k components and use Lemma 51:

$$\begin{aligned}
& \left\|[\boldsymbol{\mu}_{\not\perp}]_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 /2 \\
\leq & \left\|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + \left\|\boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 \\
= & \left\| \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \boldsymbol{\mu}_{0:k} \right\|_{\boldsymbol{\Sigma}_{0:k}}^2 \\
& + \left\| \left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{0:k}\right)^{-1} \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty} \right\|_{\boldsymbol{\Sigma}_{0:k}}^2 \\
= & \left\| \left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k}\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\
& + \left\| \left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k}\right)^{-1} \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty} \right\|^2 \\
\leq & \left(\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})\mu_1(\boldsymbol{A}_k)^{-1}\right)^{-2} \left\| \left(\mu_1(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \mu_n(\boldsymbol{A}_k) \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\
& + \mu_1(\boldsymbol{A}_k)^2 \mu_n(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-2} \mu_1(\boldsymbol{Z}_{0:k} \boldsymbol{Z}_{0:k}^\top) \mu_n(\boldsymbol{A}_k)^{-2} \left\|\boldsymbol{Q}_{k:\infty} \boldsymbol{\mu}_{k:\infty}\right\|^2,
\end{aligned}$$

where in the last transition we used Lemma 52 and the following observation:

$$\left\| \left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k}\right)^{-1} \right\| \leq \mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1} \boldsymbol{Z}_{0:k})^{-1} \leq \mu_1(\boldsymbol{A}_k)\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1}.$$

When it comes to the rest of the components, we write

$$
\begin{aligned}
&[\boldsymbol{\mu}_{\not\perp}]_{k:\infty} \\
=&\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\mu} \\
=&\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k} \\
=&\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\mu}_{0:k} \\
=&\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty} - \boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k},
\end{aligned}
$$

which yields

$$
\begin{aligned}
&\|[\boldsymbol{\mu}_{\not\perp}]_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}} \\
\leq&\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|^{1/2}\mu_n(\boldsymbol{A})^{-1}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\| \\
+&\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|^{1/2}\mu_n(\boldsymbol{A}_k)^{-1}\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{1/2}\left\|\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| \\
\leq&\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}} + \|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|^{1/2}\mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\| \\
+&\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|^{1/2}\frac{\mu_n(\boldsymbol{A}_k)^{-1}\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{1/2}}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mu_1(\boldsymbol{A}_k)^{-1}} \\
\times&\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|,
\end{aligned}
$$

where we used Lemma 52 and the fact that $\mu_n(\boldsymbol{A}) \geq \mu_n(\boldsymbol{A}_k)$ in the last transition. Combining everything together and using the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ yields the final bound.

G.5 $\mathbf{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})$

The quantity $\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})$ is exactly the variance term from Tsigler and Bartlett (2023): as for the bias term, plug in $\boldsymbol{Q}$ instead of their X, $\boldsymbol{\Sigma}$ instead of their Σ, and $\boldsymbol{A}$ instead of their A. As before, we could in principle use their algebraic decomposition from their Lemma 27, but we provide our own derivation because we want to obtain the bound in a slightly different form.

$$
\begin{aligned}
&\mathrm{tr}(\boldsymbol{A}^{-1}[\boldsymbol{Q}_{0:k}, \boldsymbol{Q}_{k:\infty}]\boldsymbol{\Sigma}[\boldsymbol{Q}_{0:k}, \boldsymbol{Q}_{k:\infty}]^\top\boldsymbol{A}^{-1}) \\
=&\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\Sigma}_{0:k}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}) + \mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}) \\
=&\mathrm{tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}\left(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\right) \\
+&\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}) \\
=&\mathrm{tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\right) \\
+&\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1})
\end{aligned}
$$

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: $\mathrm{tr}(\boldsymbol{M}_1\boldsymbol{M}_2) \leq \|\boldsymbol{M}_1\|\mathrm{tr}(\boldsymbol{M}_2)$. We use this to bound the first term as follows:

$$\begin{aligned}
&\mathrm{tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\right)\\
\leq&\mu_n(\boldsymbol{A}_k)^{-2}\mathrm{tr}\left(\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\right)\\
=&\mu_n(\boldsymbol{A}_k)^{-2}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k}\right)\\
\leq&\mu_n(\boldsymbol{A}_k)^{-2}\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\right)\\
\leq&\frac{\mu_n(\boldsymbol{A}_k)^{-2}\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2\mu_1(\boldsymbol{A}_k)^{-2}}\mathrm{tr}\left(\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right),
\end{aligned}$$

where we used Lemma 52 in the last transition.

When it comes to the second term, we simply write

$$\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}) \leq \mu_n(\boldsymbol{A}_k)^{-2}\mathrm{tr}\left(\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\right).$$

G.6 $\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\|$

In Tsigler and Bartlett (2023) 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 $\mathrm{tr}(\boldsymbol{M})$, with deviations being composed of a sub-Gaussian tail controlled by $\|\boldsymbol{M}\|_F$ and sub-exponential tail controlled by $\|\boldsymbol{M}\|$. In Tsigler and Bartlett (2023) the latter two quantities were bounded as $\|\boldsymbol{M}\|_F^2 \leq \mathrm{tr}(\boldsymbol{M})^2$ and $\|\boldsymbol{M}\| \leq \mathrm{tr}(\boldsymbol{M})$, 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 $\|\boldsymbol{M}\|_F^2 \leq \|\boldsymbol{M}\|\mathrm{tr}(\boldsymbol{M})$. This section shows the following:

$$\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\| \leq \frac{\mu_1(\boldsymbol{A}_k^{-1})^2}{\mu_n(\boldsymbol{A}_k^{-1})^2}\frac{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2} + \frac{\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|}{\mu_n(\boldsymbol{A}_k)^2}.$$

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:

$$\begin{aligned}
&\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\|\\
\leq&\|\boldsymbol{A}^{-1}\|\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|\|\boldsymbol{A}^{-1}\|\\
=&\frac{\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|}{\mu_n(\boldsymbol{A}_k)^2}.
\end{aligned}$$

For the first term we use Equation (46) to write

$$\begin{aligned}
&\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\Sigma}_{0:k}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}\| \\
=&\left\|\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}\left(\boldsymbol{I}_k+\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}\right)^{-1}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\right\| \\
=&\|\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\| \\
\le&\left\|\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\right\|\|\boldsymbol{Z}_{0:k}\boldsymbol{Z}_{0:k}^\top\|\|\boldsymbol{A}_k^{-1}\|^2 \\
\le&\left(\lambda_1^2\wedge\left\|\left(\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\right\|\right)\|\boldsymbol{Z}_{0:k}\boldsymbol{Z}_{0:k}^\top\|\|\boldsymbol{A}_k^{-1}\|^2 \\
\le&\frac{\mu_1(\boldsymbol{A}_k)^2}{\mu_n(\boldsymbol{A}_k)^2}\frac{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2}\wedge\frac{\lambda_1^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_n(\boldsymbol{A}_k)^2}.
\end{aligned}$$

Appendix H. Randomness in covariates

Lemma 53 (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. $$c\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\fallingdotseq}\ge\frac{\Lambda}{n}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|+\|\boldsymbol{\mu}_{k:\infty}\|^2\ge\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\fallingdotseq}/c;$$

3. $$\|\boldsymbol{\mu}_{\fallingdotseq}\|_{\boldsymbol{\Sigma}}^2/c\le\Lambda^2n^{-2}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2+\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2;$$

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.

$$\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\| \leq c\left(\frac{1}{n} \wedge \frac{n\lambda_1^2}{\Lambda^2} + \frac{n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2}{\Lambda^2}\right).$$

Proof Recall that on $\mathscr{A}_k(L)$ we have $\Lambda/L \leq \mu_n(\boldsymbol{A}_k) \leq \mu_1(\boldsymbol{A}_k) \leq L\Lambda$.

The proof is rather straightforward: we plug the bounds from the definition of the event $\mathscr{B}_k(c_B)$ from Section 3.2, and the bounds on eigenvalues of $\boldsymbol{A}_k$ from the definition of the event $\mathscr{A}_k(L)$ into the result of Lemma 50. 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.

$$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{\nu}\| \leq& \frac{\mu_1(\boldsymbol{A}_k)}{\mu_n(\boldsymbol{A}_k)}\frac{\sqrt{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}\left\|\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| \\ &+\mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\| \\ \leq& \frac{c_B^{3/2}L^2}{\sqrt{n}}\cdot c_B L\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| + c_B^{1/2}L\Lambda^{-1}\sqrt{n}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}.\end{aligned}$$

Note that in the last transition we used that for a positive scalar $a < 1$ we can write

$$\left\|\left(a\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| \leq a^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|.$$

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 $b \geq 1$:

$$\left\|\left(b\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| \geq b^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|.$$

2. We start with the lower bound on $\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\perp}$:

$$\begin{aligned}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\perp} \geq& \frac{1}{2}\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\left\|\left(\mu_n(\boldsymbol{A}_k)\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\ &+\|\boldsymbol{\mu}_{k:\infty}\|^2 - 9\mu_n(\boldsymbol{A}_k)^{-1}\|\boldsymbol{Q}_{0:k}^\top\boldsymbol{\mu}_{k:\infty}\|^2 \\ \geq& \frac{\Lambda}{2Lc_B n}\cdot\frac{1}{c_B L}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| \\ &+\|\boldsymbol{\mu}_{k:\infty}\|^2 - 9Lc_B\Lambda^{-1}n\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\ \geq& \frac{\Lambda}{2L^2c_B^2 n}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\| + \|\boldsymbol{\mu}_{k:\infty}\|^2(1 - 9Lc_B n\lambda_{k+1}\Lambda^{-1}),\end{aligned}$$

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 $c > 18c_B L$ in the end, then $1 - 9Lc_B n\lambda_{k+1}\Lambda^{-1} > 0.5$ since we assumed that $\Lambda > cn\lambda_{k+1}$.

Now, we do the upper bound:

$$
\begin{aligned}
&\boldsymbol{\mu}^\top \boldsymbol{\mu}_{\approx} \\
\leq& 3\mu_1(\boldsymbol{A}_k)\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1} \left\| \left(\mu_k(\boldsymbol{Z}_{0:k}^\top \boldsymbol{Z}_{0:k})^{-1}\mu_1(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k} \right\|^2 \\
&+ \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2\|\boldsymbol{A}_k^{-1/2}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \\
\leq& 3Lc_B\Lambda n^{-1} \cdot c_B L \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k} \right\|^2 \\
&+ \|\boldsymbol{\mu}_{k:\infty}\|^2 + 2Lc_B\Lambda^{-1} n \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
\leq& 3L^2 c_B^2 \Lambda n^{-1} \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k} \right\|^2 + \|\boldsymbol{\mu}_{k:\infty}\|^2 (1 + 2Lc_B\lambda_{k+1}\Lambda^{-1} n),
\end{aligned}
$$

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 $\|\boldsymbol{\mu}_{\approx}\|_{\boldsymbol{\Sigma}}^2$, is very similar to the bound on the bias term in Tsigler and Bartlett (2023), but has a slightly different form. We derive it below.

$$
\begin{aligned}
&\|\boldsymbol{\mu}_{\approx}\|_{\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} \|\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 \\
\leq& 2c_B^2 L^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 \\
&+ 2L^4 c_B^3 n^{-1} \cdot c_B n \, \|\boldsymbol{\mu}_{k:\infty}\|_{\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_B n \, \|\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^4 c_B^3 n^{-1} \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 \\
=& \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^4 L^4 \Lambda^2 n^{-2} + 3c_B^6 L^6 n^{-1} \left(n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2\right)\right) \\
&+ \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \cdot \left(2L^4 c_B^4 + 3 + 3c_B^2 L^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

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

Plugging this inequality in and taking c large enough depending on c_B and L gives the bound.

4. The upper bound on $\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})$, is also very similar to the one derived in Tsigler and Bartlett (2023), where it is exactly the the variance term, but has a slightly different form. We derive it below:

$$\begin{aligned}
&\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1})\\
\le&\frac{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2\mu_n(\boldsymbol{A}_k)^2}\mathrm{tr}\left(\left(\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_n(\boldsymbol{A}_k)\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right)\\
+&\mu_n(\boldsymbol{A}_k)^{-2}\mathrm{tr}(\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top)\\
\le&L^4c_B^3n^{-1}\cdot c_BL\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right)+L^2c_B\Lambda^{-2}n\sum_{i>k}\lambda_i^2.
\end{aligned}$$

5.

$$\begin{aligned}
&\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\|\\
\le&\frac{\mu_1(\boldsymbol{A}_k)^2}{\mu_n(\boldsymbol{A}_k)^2}\frac{\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2}\wedge\frac{\lambda_1^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_n(\boldsymbol{A}_k)^2}+\frac{\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\|}{\mu_n(\boldsymbol{A}_k)^2}\\
\le&\frac{L^4c_B^3}{n}\wedge\frac{\lambda_1^2c_Bn}{L^{-2}\Lambda^2}+c_BL^2\Lambda^{-2}\left(n\lambda_{k+1}^2+\sum_{i>k}\lambda_i^2\right).
\end{aligned}$$

In all the cases above we see that taking c large enough depending on c_B and L yields the result. ■

Appendix I. Proof of the main lower bound

First of all, we combine Lemma 49 with Lemma 53 to obtain high probability bounds on all the terms that appear in quantities of interest. The result is given by the following

Lemma 54 (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 $1 - c_y e^{-t^2/2}$ over the draw of $(\boldsymbol{y}, \hat{\boldsymbol{y}})$ all the following hold:

1. $$|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| \vee |\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}}| \leq ct\lozenge,$$
2. $$|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}| \leq ct\sigma_\eta\lozenge,$$
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. $$cM \geq \boldsymbol{\mu}^\top \boldsymbol{\mu}_{\perp} \geq c^{-1}M,$$
9. $$\|\boldsymbol{\mu}_{\perp}\|_{\boldsymbol{\Sigma}} \leq c\Lambda\lozenge/\sqrt{n},$$

Proof Parts 1, 2, 6 and 7 can be obtained directly from the corresponding parts of Lemma 49 by plugging in the bounds from Lemma 53. Parts 8 and 9 are exactly parts 2 and 3 of Lemma 53. 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 49. 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 49 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 $1 - c_1 e^{-t^2/2}$ 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.

$$\begin{aligned}\boldsymbol{y}^\top \boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \geq & (n - n\eta - c_1 t\sigma_\eta\sqrt{n} - k)\mu_1(\boldsymbol{A}_k)^{-1}\\ & -(n\eta + c_1 t\sigma_\eta\sqrt{n} + c_1 t\sqrt{n} + c_1 t^2)\|\boldsymbol{A}^{-1}\|\\ \geq & L^{-1}\Lambda^{-1}n(1 - 1/c - c_1\sigma_\eta/c - 1/c - L^2/c - c_1 L^2\sigma_\eta/c - L^2 c_1/c)\\ \geq & c^{-1}n\Lambda^{-1}.\end{aligned}$$

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 8 (Main lower bound) *For any $c_B > 0$ and $L > 1$, 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 $S > 0$:

$$S\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} \geq c^{-1}N - ct\Diamond, \tag{15}$$

$$S\|\boldsymbol{w}_{ridge}\|_{\boldsymbol{\Sigma}} \leq c\left(\left[1 + N\sigma_\eta\right]\sqrt{V + t^2\Delta V} + \Diamond\sqrt{n}\right). \tag{16}$$

That is, if $N > 2c^2 t\Diamond$, 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 54 into quantities of interest, the formulas for which are given by Lemma 35. 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 $S = (1 + \boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\lessapprox}$, and in Lemma 54 we bound $\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\lessapprox}$ and $\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}$ from below by strictly positive quantities.

Let's plug in the bounds Lemma 54 in the formulas from Lemma 35: denote the constant from Lemma 54 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

$$\begin{aligned}
&\boldsymbol{y}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\lessapprox} + (1 + \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\\
=&\boldsymbol{y}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\lessapprox} + \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} + \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}(\boldsymbol{y}_C - \Delta\boldsymbol{y}/2)\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}(\boldsymbol{y}_C + \Delta\boldsymbol{y}/2)\\
=&\boldsymbol{y}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\lessapprox} + \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} - (\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\Delta\boldsymbol{y})^2/4 + (\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y}_C)^2\\
\geq&\boldsymbol{y}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\lessapprox} + \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} - (\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\Delta\boldsymbol{y})^2/4\\
\geq&\frac{n}{c_1^2\Lambda}M - c_1 t\lozenge - c_1^2 t^2\sigma_\eta^2\lozenge^2\\
=&\frac{N}{c_1^2} - c_1 t\lozenge - c_1^2 t^2\sigma_\eta^2\lozenge^2.
\end{aligned}$$

   Recall that by Lemma 6 we have $N \geq n\lozenge^2$, which yields

$$\frac{N}{c_1^2} - c_1^2 t^2\sigma_\eta^2\lozenge^2 \geq \frac{N}{c_1^2} - \frac{c_1^2}{c^2}n\lozenge^2 \geq \frac{N}{2c_1^2},$$

   where the last transition is correct if c is taken large enough depending on c_1. Thus, we get

$$\boldsymbol{y}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\lessapprox} + (1 + \boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y})\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\hat{\boldsymbol{y}} \geq \frac{N}{2c_1^2} - c_1 t\lozenge.$$

2. $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, the first term:

$$\begin{aligned}
&\left[(1 + |\boldsymbol{\nu}^\top\boldsymbol{A}^{-1}\boldsymbol{y}|)^2 + \boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y}\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\lessapprox}\right]\|\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\Delta\boldsymbol{y}\|_{\boldsymbol{\Sigma}}\\
\leq&\left[(1 + c_1 t\lozenge)^2 + c_1^2 Mn\Lambda^{-1}\right]\sqrt{c_1\sigma_\eta^2(V + t^2\Delta V)}\\
\leq&c_1^{2.5}\left[(1 + t\lozenge)^2 + n\Lambda^{-1}M\right]\sigma_\eta\sqrt{V + t^2\Delta V}.
\end{aligned}$$

3. $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, the second term

$$\begin{aligned}
&\left[(1+|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|)(1+|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}|)+|\Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|\boldsymbol{\mu}^\top\boldsymbol{\mu}_{\perp}\right]\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}\\
\leq&\left[(1+c_1t\lozenge)(1+c_1\sigma_\eta t\lozenge)+c_1^2M\sigma_\eta n\Lambda^{-1}\right]\sqrt{c_1(V+t^2\Delta V)}\\
\leq& c_1^{2.5}\left[1+(1+\sigma_\eta)t\lozenge+\sigma_\eta t^2\lozenge^2+n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V+t^2\Delta V}\\
\leq& 2c_1^3\left[1+t\lozenge+n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V+t^2\Delta V},
\end{aligned}$$

where we used that $\sigma_\eta<1$ and $t^2\lozenge^2<n\lozenge^2<n\Lambda^{-1}M$ in the last transition (by Lemma 6).

4. $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$, the third term

$$\begin{aligned}
&\left[\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}+(1+|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|)|\Delta\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|+\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y}|\boldsymbol{\nu}^\top \boldsymbol{A}^{-1}\Delta\boldsymbol{y}|\right]\|\boldsymbol{\mu}_{\perp}\|_{\boldsymbol{\Sigma}}\\
\leq&\left[\frac{c_1n}{\Lambda}+(1+c_1t\lozenge)\frac{c_1n\sigma_\eta}{\Lambda}+\frac{c_1^2nt\sigma_\eta\lozenge}{\Lambda}\right]\frac{c_1\Lambda}{\sqrt{n}}\lozenge\\
\leq& 2c_1^3\sqrt{n}(1+t\sigma_\eta\lozenge)\lozenge,
\end{aligned}$$

where we used $c_1n\Lambda^{-1}(1+\sigma_\eta)\leq 2c_1n\Lambda^{-1}$ 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 $t\leq\sqrt{n}/c_2$ and $\eta<1/c_2$

$$\begin{aligned}
S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}/c_2\leq&\left[(1+t\lozenge)^2+n\Lambda^{-1}M\right]\sigma_\eta\sqrt{V+t^2\Delta V}\\
&+\left[1+t\lozenge+n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V+t^2\Delta V}\\
&+\sqrt{n}(1+t\sigma_\eta\lozenge)\lozenge\\
=&\lozenge^2\cdot t\sigma_\eta(\sqrt{n}+t\sqrt{V+t^2\Delta V})\\
&+\lozenge\cdot\left(\sqrt{n}+t(1+2\sigma_\eta)\sqrt{V+t^2\Delta V}\right)\\
&+\left[1+\sigma_\eta+2n\Lambda^{-1}M\sigma_\eta\right]\sqrt{V+t^2\Delta V}
\end{aligned}$$

By Lemma 6

$$V\leq 2,\quad \Delta V\leq 3/n,\quad t^2\Delta V\leq 3.$$

This allows us to obtain the final bound on $S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$: plug in the following inequalities:

$$\begin{aligned}
\sqrt{n}+t\sqrt{V+t^2\Delta V}\leq&(1+\sqrt{5})\sqrt{n},\\
\sqrt{n}+t(1+2\sigma_\eta)\sqrt{V+t^2\Delta V}\leq&(1+3\sqrt{5})\sqrt{n},\\
1+\sigma_\eta\leq&2.
\end{aligned}$$

We get for some c_3 that only depends on L, c_B:

$$S\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}} \le c_3 \left(\left[1 + n\Lambda^{-1} M \sigma_\eta\right] \sqrt{V + t^2 \Delta V} + \Diamond \sqrt{n} + t\sigma_\eta \Diamond^2 \sqrt{n} \right). \tag{51}$$

Finally, note that by Lemma 6 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.

■

Appendix J. 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 35. 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

$$\bar{\boldsymbol{Q}} := [\boldsymbol{Q}_{0:k}, \varepsilon_q \boldsymbol{Q}_{k:\infty}], \tag{52}$$

$$\bar{\boldsymbol{y}} := \varepsilon_y \boldsymbol{y}, \tag{53}$$

$$\bar{\boldsymbol{w}}_{\text{ridge}} := (\bar{\boldsymbol{Q}} + \bar{\boldsymbol{y}}\boldsymbol{\mu}^\top)^\top (\underbrace{\bar{\boldsymbol{Q}}\bar{\boldsymbol{Q}}^\top + \lambda \boldsymbol{I}_n}_{=\boldsymbol{A}})^{-1} \bar{\boldsymbol{y}}. \tag{54}$$

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 53 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 35

$$\begin{aligned}
\bar{S} :=& (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}})^2 + \boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\perp} \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}}, \\
\bar{S} \bar{\boldsymbol{w}}_{\text{ridge}} =& (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}}) \bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}} \bar{\boldsymbol{\mu}}_{\perp}, \\
\bar{S} \boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}} =& \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}} \boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\perp} + (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}}) \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1} \bar{\boldsymbol{y}},
\end{aligned}$$

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 J.1 we bound $\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}$ from above, in Section J.2 we bound $\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}$ from below. In Section J.3 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.

J.1 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 55 *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 3.2.

1. Using the expressions that we derived in Section G.2 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 52 in the penultimate line.

2. Use Sherman-Morrison-Woodbury for the matrix $\boldsymbol{A}^{-1}$ (Equation (45)):

$$\boldsymbol{A}^{-1} = (\boldsymbol{A}_k + \boldsymbol{Q}_{0:k}\boldsymbol{Q}_{0:k}^\top)^{-1} = \boldsymbol{A}_k^{-1} - \boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k}(\boldsymbol{I}_k + \boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Q}_{0:k})^{-1}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}_k^{-1}.$$

   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 (45)) 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 39) for some absolute constant c_1 and any $s > 0$, the probability of the following event is at most $2\exp(-s/c_1)$

$$|\|\boldsymbol{P}_{0:k}^{\perp}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 - \|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}}^2\mathrm{tr}(\boldsymbol{P}_{0:k}^{\perp})| > \sigma_x^2\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}}^2\max(\sqrt{s}\|\boldsymbol{P}_{0:k}^{\perp}\|_F, s\|\boldsymbol{P}_{0:k}^{\perp}\|).$$

Once again, $\boldsymbol{P}_{0:k}^{\perp}$ is a projector of rank $n-k$, so

$$\mathrm{tr}(\boldsymbol{P}_{0:k}^{\perp}) = n-k > n/2, \quad \|\boldsymbol{P}_{0:k}^{\perp}\|_F = \sqrt{n-k} \leq \sqrt{n}, \quad \|\boldsymbol{P}_{0:k}^{\perp}\|_F = 1.$$

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 $\bar{S}\boldsymbol{\mu}^{\top}\bar{\boldsymbol{w}}_{\mathrm{ridge}}$ is given by the following lemma.

Lemma 56 *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* $n\Lambda^{-1}M \geq a^{-1}\lozenge$, *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* $n\Lambda^{-1}M < a^{-1}\lozenge$, *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

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

First of all, denote the constant from Lemma 53 as c_1. By that lemma on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$ we have

$$\begin{aligned}
\bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\perp\!\!\!\perp} \leq& \mu_n(\boldsymbol{A})^{-1}\|\bar{\boldsymbol{y}}\|^2 \boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\perp\!\!\!\perp} \\
\leq& L\Lambda^{-1}n \cdot c_1 M; \\
\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\| \leq& c_1 \lozenge.
\end{aligned}$$

Recall that we indeed can apply Lemma 53 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 49 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

$$|\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}| = |\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| \leq c_1 t \lozenge.$$

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,

$$\begin{aligned}
\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}} =& \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\boldsymbol{\mu}^\top (\boldsymbol{I}_p - \bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{Q}}) + (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}} \\
\leq& c_1^2 n\Lambda^{-1}M + c_1 t\lozenge + (c_1 t\lozenge)^2 \\
\leq& c_1^2 n\Lambda^{-1}M + c_1 t\lozenge + c_1 n\lozenge^2 \\
\leq& n\Lambda^{-1}M(c_1^2 + c_1) + c_1 t a n\Lambda^{-1}M,
\end{aligned}$$

where we used that $t < \sqrt{n}$, $\lozenge^2 \leq \Lambda^{-1}M$ (Lemma 6) and $\lozenge < an\Lambda^{-1}M$ 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 55, by which there exists a constant c_2 that only depends on L, σ_x, c_B such that on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B) \cap \mathscr{C}$

$$\begin{aligned}\|\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}\|^2 \geq& c_2^{-1} n^{-1} \left\| \left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2, \\ \|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \geq& c_2^{-1}\Lambda^{-2} n \|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}.\end{aligned}$$

Now we use the same expressions as we derived in Section G.2 to write

$$\begin{aligned}\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\|^2 =& \|\boldsymbol{A}^{-1}(\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k} + \varepsilon_q \boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty})\|^2 \\ =& \left\| \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{Z}_{0:k}^\top \boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 \\ &+ \|\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\|^2 \\ &+ 2\varepsilon_q (\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty})^\top \boldsymbol{A}^{-2}(\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}) \\ \geq& c_2^{-1}\lozenge^2/2 + 2\varepsilon_q (\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty})^\top \boldsymbol{A}^{-2}(\boldsymbol{Q}_{0:k}\boldsymbol{\mu}_{0:k}).\end{aligned}$$

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 $\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\| \geq (2c_2)^{-1/2}\lozenge$. That statement doesn't involve $\boldsymbol{y}$, so $\boldsymbol{y}$ is still independent of this event. Thus, conditionally on it, by Lemma 43 for an absolute constant $c_{y,2}$ with probability at least $c_{y,2}^{-1}$ over the choice of $\boldsymbol{y}$ we have $|\boldsymbol{y}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}| \geq c_{y,2}^{-1}\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\|$. 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 $\boldsymbol{y}$ $|\boldsymbol{y}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}| \leq c_1\sqrt{n}\lozenge$. 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 $(\varepsilon_q, \varepsilon_y)$

$$\begin{aligned}\bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}} \leq& -c_{y,2}^{-1}(2c_2)^{-1/2}\lozenge, \\ \bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}} \leq& n\Lambda^{-1}M(c_1^2 + 2c_1) - c_{y,2}^{-1}(2c_2)^{-1/2}\lozenge \\ <& 0,\end{aligned}$$

where the last transition holds for a large enough depending on $c_1, c_2, c_{y,1}, c_{y,2}$ since $n\Lambda^{-1}M < a^{-1}\lozenge$. ∎

J.2 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 57 *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

$$\|\bar{S}\bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}^2 \geq 0.5\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\bar{\boldsymbol{\mu}}_{\perp\!\!\!\perp}\|_{\boldsymbol{\Sigma}}^2 - 7(\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})^2\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\|_{\boldsymbol{\Sigma}}^2.$$

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:

$$\begin{aligned}&\left\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}\bar{\boldsymbol{\mu}}_{\perp\!\!\!\perp}\right\|_{\boldsymbol{\Sigma}}^2 \\ =& \left\|\bar{\boldsymbol{Q}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 + (\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2\left\|\bar{\boldsymbol{\mu}}_{\perp\!\!\!\perp}\right\|_{\boldsymbol{\Sigma}}^2 + 2\varepsilon_y f_1(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}, \varepsilon_q \boldsymbol{y}),\end{aligned}$$

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}&\left\|\bar{\boldsymbol{\mu}}_{\perp\!\!\!\perp}\right\|_{\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}) \\ \geq& \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}) \\ \geq& \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 - 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

$$\{\varepsilon_q f_2(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}) \geq 0, \varepsilon_y f_1(\boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{\mu}, \varepsilon_q \boldsymbol{y}) \geq 0, \}$$

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 58 *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:*

$$\begin{aligned}
\left\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 \geq& c_y^{-1} tr(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}),\\
\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \geq& (n-k)\mu_1(\boldsymbol{A}_k)^{-1} - c_y(t\sqrt{n} + t^2)\|\boldsymbol{A}^{-1}\|,\\
\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} \leq& n\|\boldsymbol{A}^{-1}\|,\\
\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \leq& c_y(tr(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}) + t^2\|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\|),\\
|\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\boldsymbol{y}| \leq& c_y t\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\|.
\end{aligned}$$

Proof The first inequality is a direct application of Lemma 44, and the remaining were shown as a part of Lemma 49. Note that only the last inequality depends on ε_q, and formally Lemma 49 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 49 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 59 *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.

$$\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) = \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\Sigma}_{0:k}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}) + \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1})$$

For the first term we use the same formula as in Section G.5:

$$\begin{aligned}
&\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k}\boldsymbol{\Sigma}_{0:k}\boldsymbol{Q}_{0:k}^\top\boldsymbol{A}^{-1}) \\
=&\operatorname{tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\right) \\
\geq&\mu_1(\boldsymbol{A}_k)^{-2}\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})\operatorname{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{Z}_{0:k}^\top\boldsymbol{A}_k^{-1}\boldsymbol{Z}_{0:k}\right)^{-2}\right) \\
\geq&\frac{\mu_n(\boldsymbol{A}_k)^2\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})}{\mu_1(\boldsymbol{A}_k)^2\mu_1(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^2}\operatorname{tr}\left(\left(\mu_k(\boldsymbol{Z}_{0:k}^\top\boldsymbol{Z}_{0:k})^{-1}\mu_1(\boldsymbol{A}_k)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right) \\
\geq&L^4c_B^3n^{-1}\cdot c_B^2L^2\operatorname{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-2}\right),
\end{aligned}$$

where we used Lemma 52 in the penultimate line and the definition of the event $\mathscr{B}_k(c_B)$ from Section 3.2 in the last transition.

When it comes to the second term, we once again (as in the proof of Lemma 59) 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

$$\begin{aligned}
&\operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}) \\
\geq&\mu_1(\boldsymbol{A}_k)^{-2}\operatorname{tr}(\boldsymbol{A}_k\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{A}_k) \\
\geq&\mu_1(\boldsymbol{A}_k)^{-2}\operatorname{tr}(\boldsymbol{P}_{0:k}^\perp\boldsymbol{A}_k\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{A}_k\boldsymbol{P}_{0:k}^\perp) \\
=&\mu_1(\boldsymbol{A}_k)^{-2}\operatorname{tr}(\boldsymbol{P}_{0:k}^\perp\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{P}_{0:k}^\perp) \\
=&\mu_1(\boldsymbol{A}_k)^{-2}\sum_{i>k}\lambda_i^2\boldsymbol{z}_i^\top\boldsymbol{P}_{0:k}^\perp\boldsymbol{z}_i,
\end{aligned}$$

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 59, by Hanson-Wright inequality (Lemma 39) 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 $1 - c_2e^{-n/c_2}$

$$\boldsymbol{z}_i^\top\boldsymbol{P}_{0:k}^\perp\boldsymbol{z}_i \geq n/c_2.$$

By Lemma 9 from Bartlett et al. (2020), 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 51 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 52 in the penultimate line and the definition of the event $\mathscr{B}_k(c_B)$ from Section 3.2 in the last transition.

■

Finally, we can put everything together in the following

Lemma 60 *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*

$$\Lambda > c \left(n\lambda_{k+1} + \sqrt{n \sum_{i>k} \lambda_i^2} \right), \tag{55}$$

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 59, 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 53 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 53 on $\mathscr{A}_k(L) \cap \mathscr{B}_k(c_B)$

$$\begin{aligned}\|\boldsymbol{A}^{-1}\bar{\boldsymbol{\nu}}\| \leq& c_2\lozenge,\\ \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}) \leq& c_2 V\\ \|\boldsymbol{A}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\| \leq& c_2\Delta V.\end{aligned}$$

Now denote the constant from Lemma 58 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:

$$\begin{aligned}\left\|\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 \geq& c_y^{-1}c_1^{-1}V,\\ \boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y} \geq& (n-k)\mu_1(\boldsymbol{A}_k)^{-1} - c_y(t\sqrt{n}+t^2)\|\boldsymbol{A}^{-1}\|\\ \geq& n\Lambda^{-1}(L^{-1}(1-k/n) - c_y L\Lambda^{-1}(t\sqrt{n}+t^2),\\ \boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{y} \leq& nL\Lambda^{-1},\\ \|\boldsymbol{Q}^\top\boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \leq& c_y c_2(V + t^2\Delta V),\\ |\bar{\boldsymbol{\nu}}^\top\boldsymbol{A}^{-1}\boldsymbol{y}| \leq& c_y c_2 t\lozenge.\end{aligned}$$

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

$$\begin{aligned}&\left\|\boldsymbol{Q}_{k:\infty}^\top\boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\\ \leq& \left\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\Sigma}_{k:\infty}\boldsymbol{Q}_{k:\infty}^\top\right\|\mu_n(\boldsymbol{A})^{-2}\left\|\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|^2\\ \leq& c_B\left(\sum_{i>k}\lambda_i^2 + n\lambda_{k+1}^2\right)\cdot L^2\Lambda^{-2}\cdot c_B n\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\\ \leq& \frac{c_B^2L^2}{c^2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2,\end{aligned}$$

where in the last line we used the assumption from Equation (55).

Combining all that with Lemma 57 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 $(\varepsilon_y, \varepsilon_q)$

$$\begin{aligned}
\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}^2 \geq & \frac{1}{2}\left\|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\right\|_{\boldsymbol{\Sigma}}^2 \\
& + \frac{1}{2}(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 \left(\left\|(\boldsymbol{I}_k - \boldsymbol{Q}_{0:k}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{0:k})\boldsymbol{\mu}_{0:k}\right\|_{\boldsymbol{\Sigma}_{0:k}}^2 + 0.5\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\right) \\
& - \frac{7}{2}(\boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 \left\|\boldsymbol{Q}_{k:\infty}^\top \boldsymbol{A}^{-1}\boldsymbol{Q}_{k:\infty}\boldsymbol{\mu}_{k:\infty}\right\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
& - 7(\bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\boldsymbol{y})^2 \|\boldsymbol{Q}^\top \boldsymbol{A}^{-1}\boldsymbol{y}\|_{\boldsymbol{\Sigma}}^2 \\
\geq & \frac{1}{2c_1 c_y} V \\
& + \frac{1}{2c_3^2} n^2 \Lambda^{-2} \cdot \left(c_1^{-1}\Lambda^2 n^{-2}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + 0.5\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\right) \\
& - \frac{7}{2} n^2 \Lambda^{-2} \cdot \frac{c_B^2 L^2}{c^2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 - 7c_y^2 c_2^2 t^2 \lozenge^2 \cdot c_y c_2 (V + t^2 \Delta V).
\end{aligned}$$

If c is large enough, namely if $7c_B^2 L^2 c_3^2 < 0.25c^2$, then we have

$$\begin{aligned}
& \frac{1}{2c_3^2} n^2 \Lambda^{-2} \cdot \left(c_1^{-1}\Lambda^2 n^{-2}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + 0.5\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\right) \\
& - \frac{7}{2} n^2 \Lambda^{-2} \cdot \frac{c_B^2 L^2}{c^2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
\geq & \frac{1}{2c_3^2} \cdot \left(c_1^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + 0.25 n^2 \Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\right) \\
\geq & \frac{\min(c_1^{-1}, 0.25)}{2c_3^2} n \lozenge^2 / 2.
\end{aligned}$$

That is, for a large enough constant c_4 that only depends on σ_x, c_B, L, we have

$$\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}^2 \geq c_4^{-1}(V + n\lozenge^2) - c_4 t^2 \lozenge^2 (V + t^2 \Delta V).$$

By Lemma 6 $V \leq 2$ and $t^2 \Delta V \leq 3t^2/n \leq 3$. Thus,

$$\|\bar{S}\boldsymbol{\mu}^\top \bar{\boldsymbol{w}}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}^2 \geq c_4^{-1}(V + n\lozenge^2(1 - 5c_4^2 t^2/n)) \geq c_5^{-1}(V + n\lozenge^2),$$

provided that c_5 is a large enough constant depending on c_4, and $t = \sqrt{n/c_5}$.

■

J.3 The ratio

Finally we can put the bound on the numerator together with the bound on the denominator and obtain the following

Theorem 9 (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 $\boldsymbol{\mu}^\top \boldsymbol{w}_{ridge} < 0$ with probability at least $c_y^{-1}(\mathbb{P}(\mathscr{A}_k(L)) - ce^{-n/c})_+$, where 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

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

Proof First of all, it is enough to show the statement for $\bar{\boldsymbol{w}}_{\text{ridge}}$ (as defined in the beginning of Section J) instead of $\boldsymbol{w}_{\text{ridge}}$ as it has the same distribution.

The straightforward combination of Lemmas 60 and 56 almost does the job, but we also need to show that $\bar{S} > 0$ with high probability.

Recall that

$$\bar{S} = (1 + \bar{\boldsymbol{\nu}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}})^2 + \boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\perp} \bar{\boldsymbol{y}}^\top \boldsymbol{A}^{-1}\bar{\boldsymbol{y}}.$$

By Lemma 53, 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

$$\boldsymbol{\mu}^\top \bar{\boldsymbol{\mu}}_{\perp} \geq M/c > 0.$$

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 3.2). Now the first part of the theorem is a direct consequence of part 2 of Lemma 56, while the second part of the theorem is a direct combination of Lemma 60 with the first part of 56. ∎

Theorem 11 (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,

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

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

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

Proof First of all, denote the constants from Theorem 9 as $a_u, c_u, c_{y,u}$ (here index u stands for "upper bound"). Note that by that theorem, regardless whether $n\Lambda^{-1}M < a^{-1}\lozenge$ or $n\Lambda^{-1}M \ge a^{-1}\lozenge$ it still holds for any $t_u \in (0, \sqrt{n}/c_{y,u})$ that the probability of the event

$$\left\{\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}}{\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}} \le c_u(1+t_u)\frac{n\Lambda^{-1}M}{\sqrt{V+n\lozenge^2}}\right\}$$

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

$$\alpha_\varepsilon < c_u(1+t_u)\frac{n\Lambda^{-1}M}{\sqrt{V+n\lozenge^2}}.$$

When it comes to the lower bound, recall that by Lemma 5, 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 8 is applicable. Denote the constant from Theorem 8 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 $\boldsymbol{y}$

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{w}_{\text{ridge}}}{\|\boldsymbol{w}_{\text{ridge}}\|_{\boldsymbol{\Sigma}}} \ge c_\ell^{-2}\frac{N - c_\ell^2 t_\ell \lozenge}{\sqrt{V + t_\ell^2 \Delta V} + \lozenge\sqrt{n}} \ge c_\ell^{-2}\frac{N/2}{\sqrt{V(1+4t_\ell^2)} + \lozenge\sqrt{n}},$$

where the last transition is made under the assumption that $c_\ell^2 t_\ell \lozenge \le N$, and also uses the fact that $\Delta V \le 4V$ (Lemma 6).

Thus, if we take $a = 2c_\ell^2 t_\ell$ and $t_\ell, n, \varepsilon, \delta$ are such that

$$(1-\delta-c_B e^{-n/c_B})_+(1-c_\ell e^{-t_\ell^2/2})_+ \ge 1-\varepsilon,$$

then under the condition $n\Lambda^{-1}M \ge a\lozenge$ we get

$$\alpha_\varepsilon \ge c_\ell^{-2}\frac{n\Lambda^{-1}M/2}{\sqrt{V(1+4t_\ell^2)} + \lozenge\sqrt{n}}.$$

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

$$c_y e^{-c_1/c_u} \le c_{y,u}^{-1}/4, \quad c_y e^{-c_1/c_u} \le \frac{1}{4}, \quad c_B e^{-c_1/c_B} \le \varepsilon/4.$$

Fourth, take $\delta = \varepsilon/4$ – an absolute constant. Finally, take t_ℓ such that $c_\ell e^{-t_\ell^2/2} \le \varepsilon/2$—a constant that only depends on c_ℓ (which, in its turn, only depends on L and σ_x). Combining all gives

$$\begin{aligned}&(1-\delta-c_B e^{-n/c_B})_+(1-c_{y,\ell}e^{-t_\ell^2/2})_+\\ \ge&(1-\varepsilon/4-\varepsilon/4)_+(1-\varepsilon/2)_+\\ >&1-\varepsilon,\\ &(c_{y,u}^{-1}-c_{y,u}e^{-t_u^2/c_{y,u}}-c_y e^{-n/c_u})_+(1-\delta-c_u e^{-n/c_u})_+\\ >&(c_{y,u}^{-1}-c_{y,u}^{-1}/2-c_{y,u}^{-1}/4)(1-1/4-1/4)\\ =&c_{y_u}^{-1}/16 \ge \varepsilon,\end{aligned}$$

which finishes the proof. ■

Appendix K. Analysis of ridge regularization

Lemma 61 *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

$$\begin{aligned}
f(\boldsymbol{w}) :=& \boldsymbol{v}^\top \boldsymbol{w}/\|\boldsymbol{w}\|, \\
\nabla_{\boldsymbol{w}} f(\boldsymbol{w}) =& \boldsymbol{v}/\|\boldsymbol{w}\| - \boldsymbol{w}\cdot \boldsymbol{v}^\top \boldsymbol{w}/\|\boldsymbol{w}\|^3, \\
\dot{\boldsymbol{w}} := \frac{d}{dt}\boldsymbol{w} =& -\boldsymbol{M}(\boldsymbol{I}_p + t\boldsymbol{M})^{-2}\boldsymbol{v} \\
=& -t^{-1}(\boldsymbol{I}_p + t\boldsymbol{M} - \boldsymbol{I}_p)(\boldsymbol{I}_p + t\boldsymbol{M})^{-2}\boldsymbol{v} \\
=& -t^{-1}(\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{v} + t^{-1}(\boldsymbol{I}_p + t\boldsymbol{M})^{-2}\boldsymbol{v} \\
=& -t^{-1}\left(\boldsymbol{w} + (\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{w}\right). \\
t\boldsymbol{v}^\top \dot{\boldsymbol{w}} =& -\boldsymbol{v}^\top \boldsymbol{w} + \|\boldsymbol{w}\|^2, \\
t\boldsymbol{w}^\top \dot{\boldsymbol{w}} =& -\|\boldsymbol{w}\|^2 + \boldsymbol{w}^\top(\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{w}, \\
t\|\boldsymbol{w}\|^3\langle \nabla f(\boldsymbol{w}), \dot{\boldsymbol{w}}\rangle =& \|\boldsymbol{w}\|^4 - \boldsymbol{v}^\top \boldsymbol{w}\cdot \boldsymbol{w}^\top(\boldsymbol{I}_p + t\boldsymbol{M})^{-1}\boldsymbol{w} \\
=& \left\langle (\boldsymbol{I}_p + t\boldsymbol{M})^{-1/2}\boldsymbol{v}, (\boldsymbol{I}_p + t\boldsymbol{M})^{-3/2}\boldsymbol{v}\right\rangle^2 \\
& - \|(\boldsymbol{I}_p + t\boldsymbol{M})^{-1/2}\boldsymbol{v}\|^2 \cdot \|(\boldsymbol{I}_p + t\boldsymbol{M})^{-3/2}\boldsymbol{v}\|^2 \\
\leq& 0,
\end{aligned}$$

where the last transition is by Cauchy-Schwartz and we used the brackets $\langle \boldsymbol{u}_1, \boldsymbol{u}_2\rangle$ to denote the scalar product $\boldsymbol{u}_1^\top \boldsymbol{u}_2$. 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 62 (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 $\lambda' > \lambda$*

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

Proof First of all, note that $\Lambda(\lambda') > \Lambda(\lambda) > n\lambda_k$. Thus, by Lemma 16 we can show that for some absolute constant $c_1 > 0$

$$\frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)} \vee \sqrt{n}\lozenge_a(\Lambda_1)} \leq c_1\left(1 + \frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)} \vee \sqrt{n}\lozenge_a(\Lambda_0)}\right),$$

where we denoted $\Lambda_0 = \Lambda(\lambda)$, $\Lambda_1 = \Lambda(\lambda')$ and used the notation from Lemma 16.

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

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

Thus, by Lemma 17, $\frac{N_a(\Lambda)}{\sqrt{n}\diamondsuit_a(\Lambda)}$ 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

$$\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\diamondsuit_a(\Lambda_0)} < \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\diamondsuit_a(\Lambda_1)}.$$

Then $\sqrt{V_a(\Lambda_0)} \geq \sqrt{n}\diamondsuit_a(\Lambda_0)$, otherwise we would have

$$\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\diamondsuit_a(\Lambda_0)} = \frac{N_a(\Lambda_0)}{\sqrt{n}\diamondsuit_a(\Lambda_0)} \geq \frac{N_a(\Lambda_1)}{\sqrt{n}\diamondsuit_a(\Lambda_1)} \geq \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\diamondsuit_a(\Lambda_1)}.$$

Moreover, if $\sqrt{V_a(\Lambda_1)} < \sqrt{n}\diamondsuit_a(\Lambda_1)$ then by Intermediate Value Theorem we can take such $\Lambda_{0.5}$ that $\sqrt{V_a(\Lambda_{0.5})} = \sqrt{n}\diamondsuit_a(\Lambda_{0.5})$, and we'll once again have

$$\frac{N_a(\Lambda_{0.5})}{\sqrt{V_a(\Lambda_{0.5})}} \wedge \frac{N_a(\Lambda_{0.5})}{\sqrt{n}\diamondsuit_a(\Lambda_{0.5})} = \frac{N_a(\Lambda_{0.5})}{\sqrt{n}\diamondsuit_a(\Lambda_{0.5})} \geq \frac{N_a(\Lambda_1)}{\sqrt{n}\diamondsuit_a(\Lambda_1)} \geq \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\diamondsuit_a(\Lambda_1)}.$$

In case $\Lambda_{0.5}$ as above exists set $\Lambda = \Lambda_{0.5}$, otherwise set $\Lambda = \Lambda_1$. Now we have

$$\Lambda_1 \geq \Lambda > \Lambda_0,$$
$$\sqrt{V_a(\Lambda)} \geq \sqrt{n}\diamondsuit_a(\Lambda), \quad \sqrt{V_a(\Lambda_0)} \geq \sqrt{n}\diamondsuit_a(\Lambda_0),$$
$$\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\diamondsuit_a(\Lambda_0)} < \frac{N_a(\Lambda)}{\sqrt{V_a(\Lambda)}} \wedge \frac{N_a(\Lambda)}{\sqrt{n}\diamondsuit_a(\Lambda)} \geq \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\diamondsuit_a(\Lambda_1)}.$$

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

$$\begin{aligned}
\frac{N_a(\Lambda)}{\sqrt{V_a(\Lambda)}} =& \frac{\sum_i \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} \\
=& \frac{\sum_{i=1}^{k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} + \frac{\sum_{i>k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} \\
\leq& \frac{\sum_{i=1}^{k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} + \frac{\sum_{i>k_0} \frac{\mu_i^2}{\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} \\
=& \frac{\sum_{i=1}^{k_0} \frac{\mu_i^2}{\lambda_i+\Lambda/n}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} + \frac{\sum_{i>k_0} \mu_i^2}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(n\lambda_i/\Lambda+1)^2}}}
\end{aligned}$$

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 17, the first multiplier is a non-increasing function of Λ. Moreover, if we plug Λ_0 instead of Λ we get

$$\frac{\sum_{i=1}^{k_0}\frac{\mu_i^2}{\lambda_i+\Lambda_0/n}}{\sqrt{\sum_{i=1}^{k_0}\frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda_0/n)^2}}} \le 2\frac{\sum_{i=1}^{k_0}\frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda_0/n)^2}}{\sqrt{\sum_{i=1}^{k_0}\frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda_0/n)^2}}} = 2\sqrt{\sum_{i=1}^{k_0}\frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda_0/n)^2}}$$
$$\le 2\sqrt{n}\diamondsuit_a(\Lambda_0) \le 2\sqrt{V_a(\Lambda_0)} \le 2\sqrt{2},$$

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 16) and $V(\Lambda_0) < 2$ (Lemma 6).

Thus, the first multiplier starts less than a constant and stays less than a constant. For the second multiplier we have

$$\frac{\sqrt{\sum_{i=1}^{k_0}\frac{\lambda_i\mu_i^2}{(\lambda_i+\Lambda/n)^2}}}{\sqrt{\sum_i \frac{\lambda_i^2/n}{(\lambda_i+\Lambda/n)^2}}} \le \frac{\sqrt{n}\diamondsuit_a(\Lambda)}{\sqrt{V_a(\Lambda)}} \le 1.$$

Overall, we've got that either

$$\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\diamondsuit_a(\Lambda_0)} \ge \frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\diamondsuit_a(\Lambda_1)},$$

or

$$\frac{N_a(\Lambda_1)}{\sqrt{V_a(\Lambda_1)}} \wedge \frac{N_a(\Lambda_1)}{\sqrt{n}\diamondsuit_a(\Lambda_1)} \le \frac{N_a(\Lambda)}{\sqrt{V_a(\Lambda)}} \le 2\sqrt{2} + 2\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} = 2\sqrt{2} + 2\frac{N_a(\Lambda_0)}{\sqrt{V_a(\Lambda_0)}} \wedge \frac{N_a(\Lambda_0)}{\sqrt{n}\diamondsuit_a(\Lambda_0)},$$

which implies the desired result.

■

Corollary 63 (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 $i \leq k$

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

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

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

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 11 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

$$\begin{aligned}
&\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\Diamond(\lambda')} \\
=&\frac{\left(1 + \lambda_i^{-1}n^{-1}\Lambda(\lambda')\right)^{-1}\lambda_i^{-1}\mu_i^2}{\sqrt{n^{-1}\left(1 + \lambda_i^{-1}n^{-1}\Lambda(\lambda')\right)^{-1}\lambda_i^{-1}\mu_i^2}} \\
=&\sqrt{n/\lambda_i}
\end{aligned}$$

—doesn't depend on λ' once again.

That is

$$\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\Diamond(\lambda')} = \frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\Diamond(\lambda)} \geq a = a_0.$$

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 11 are satisfied for both λ and λ', which yields

$$\alpha_\varepsilon(\lambda)/c_0 \leq \frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{V(\lambda)}+\sqrt{n}\Diamond(\lambda)} \leq c_0\alpha_\varepsilon(\lambda), \quad \alpha_\varepsilon(\lambda')/c_0 \leq \frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{V(\lambda')}+\sqrt{n}\Diamond(\lambda')} \leq c_0\alpha_\varepsilon(\lambda').$$

We've already seen that $\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\Diamond(\lambda')} = \frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{n}\Diamond(\lambda)}$, so the only thing we need to study is $V(\lambda)$. Namely, we are going to show that under the assumptions we made $V(\lambda) < n\Diamond^2(\lambda)$ and $V(\lambda') < n\Diamond^2(\lambda')$. That will finish the proof since in that case

$$\begin{aligned}\alpha_\varepsilon(\lambda')/c_0 \leq \frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{V(\lambda')}+\sqrt{n}\Diamond(\lambda')} &\leq \frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\Diamond(\lambda')} \\ &= \frac{n\Lambda(\lambda')^{-1}M(\lambda)}{\sqrt{n}\Diamond(\lambda)} \leq 2\frac{n\Lambda(\lambda')^{-1}M(\lambda)}{\sqrt{V(\lambda')}+\sqrt{n}\Diamond(\lambda)} \leq 2c_0\alpha_\varepsilon(\lambda),\end{aligned}$$

and analogously $\alpha_\varepsilon(\lambda)/c_0 \leq 2c_0\alpha_\varepsilon(\lambda')$.

Thus, in the rest of the proof we show that $V(\lambda) < n\Diamond^2(\lambda)$ and $V(\lambda') < n\Diamond^2(\lambda')$. 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

$$n^{-1}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1}+n\Lambda(\lambda)^{-1}\boldsymbol{I}_k\right)^{-2}\right)+n^{-1}\sum_{i>k}\lambda_i^2 \leq \|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}},$$

which holds since

$$\begin{aligned}
&n^{-1}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1}+n\Lambda(\lambda)^{-1}\boldsymbol{I}_k\right)^{-2}\right)+n^{-1}\sum_{i>k}\lambda_i^2\\
<&n^{-1}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{0:k}^{-1}\right)^{-2}\right)+n^{-1}\sum_{i>k}\lambda_i^2\\
=&\sum_i\lambda_i^2\\
\leq&n\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2.
\end{aligned}$$

We showed that $V(\lambda)\leq n\Diamond^2(\lambda)$. Note that $V(\lambda')\leq n\Diamond^2(\lambda')$ by exactly the same argument.

2. $\boldsymbol{\mu}=\boldsymbol{e}_i$ for $i\leq k$. Write out V and $\Diamond$ again:

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

By the same argument as before, since

$$\frac{n\lambda_i\mu_i^2}{(1+n\lambda_i/\Lambda(\lambda))^2}\geq\sum_i\lambda_i^2,$$

we have $V(\lambda)<n\Diamond(\lambda)^2$. When it comes to λ', we can write

$$\frac{n\lambda_i\mu_i^2}{(1+n\lambda_i/\Lambda(\lambda'))^2}\geq\frac{n\lambda_i\mu_i^2}{(1+n\lambda_i/\Lambda(\lambda))^2}\geq\sum_i\lambda_i^2,$$

which yields $V(\lambda')<n\Diamond(\lambda')^2$.

■

Lemma 64 *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,*

$$\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}} \leq C^{-1}n^{-1}\Lambda(\lambda)\|\boldsymbol{\mu}_{k:\infty}\|^2.$$

4. *Take*[6] *$\boldsymbol{\mu}_{0:k}$ which balances $\boldsymbol{\mu}_{k:\infty}$ in the following sense:*

$$nC^{-1}\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2 \geq \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} \geq n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}. \tag{29}$$

5. *Scale $\boldsymbol{\mu}$ up*[7] *if needed, so it holds that*

$$n\diamondsuit^2(\lambda) \geq V(\lambda) \quad \text{and } N(\lambda) \geq a\diamondsuit(\lambda).$$

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 11. 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 11 implies that

$$\alpha_\varepsilon(\lambda) \geq c_1^{-1}\frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{V(\lambda)} + \sqrt{n}\diamondsuit(\lambda)} \geq c_1^{-1}\frac{n\Lambda(\lambda)^{-1}M(\lambda)}{2\sqrt{n}\diamondsuit(\lambda)}.$$

At the same time, by Theorem 11 for any $\lambda' > \lambda$

$$\alpha_\varepsilon(\lambda') \leq c_1\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{V(\lambda')} + \sqrt{n}\diamondsuit(\lambda')} \leq c_1\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\diamondsuit(\lambda')}.$$

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 (29) we have

$$\begin{aligned}
n\Lambda(\lambda)^{-1}M(\lambda) \geq& n\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2,\\
n\Lambda(\hat{\lambda})^{-1}M(\hat{\lambda}) \leq& \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} + n\Lambda(\hat{\lambda})^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2\\
\leq& 2n\Lambda(\hat{\lambda})^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2,\\
n\diamondsuit(\lambda)^2 \leq& \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}} + n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\
\leq& 2\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}},\\
n\diamondsuit(\hat{\lambda})^2 \geq& \left\|\left(\Lambda(\hat{\lambda})n^{-1}\boldsymbol{\Sigma}^{-1}_{0:k} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}^{-1/2}_{0:k}\boldsymbol{\mu}_{0:k}\right\|^2\\
\geq& \frac{1}{4}\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}^{-1}_{0:k}},
\end{aligned}$$

6. Note that such $\boldsymbol{\mu}_{0:k}$ exists because of how we chose $\boldsymbol{\mu}_{k:\infty}$.
7. Note that the previous conditions were homogeneous in $\boldsymbol{\mu}$, so multiplying it by a scalar does not break them.

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

$$\alpha_\varepsilon(\lambda) \geq c_1^{-1}\frac{n\Lambda(\lambda)^{-1}M(\lambda)}{\sqrt{n}\diamondsuit(\lambda)} \geq \frac{1}{2c_1}\frac{n\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2}{\sqrt{2}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}} =$$
$$= \frac{C}{8\sqrt{2}c_1}\frac{2n\Lambda(\hat{\lambda})^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2}{\frac{1}{2}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}} \geq \frac{C}{8\sqrt{2}c_1^2}\cdot c_1\frac{n\Lambda(\hat{\lambda})^{-1}M(\hat{\lambda})}{\sqrt{n}\diamondsuit(\hat{\lambda})} \geq \frac{C}{8\sqrt{2}c_1^2}\alpha_\varepsilon(\hat{\lambda}).$$

We obtained the result for $\lambda' = \hat{\lambda}$. To extend the result for all $\lambda' > \hat{\lambda}$ note that by Lemma 16 and the fact that $N_a(\lambda)/\diamondsuit_a(\lambda)$ is a non-increasing function of λ for some absolute constant $c_2 > 1$ we can write

$$\alpha_\varepsilon(\lambda) \geq \frac{C}{8\sqrt{2}c_1^2}\cdot c_1\frac{n\Lambda(\hat{\lambda})^{-1}M(\hat{\lambda})}{\sqrt{n}\diamondsuit(\hat{\lambda})} \geq \frac{C}{8\sqrt{2}c_1^2c_2}\cdot c_1\frac{N_a(\hat{\lambda})}{\sqrt{n}\diamondsuit_a(\hat{\lambda})} \geq$$
$$\geq \frac{C}{8\sqrt{2}c_1^2c_2}\cdot c_1\frac{N_a(\lambda')}{\sqrt{n}\diamondsuit_a(\lambda')} \geq \frac{C}{8\sqrt{2}c_1^2c_2^2}\cdot c_1\frac{n\Lambda(\lambda')^{-1}M(\lambda')}{\sqrt{n}\diamondsuit(\lambda')} \geq \frac{C}{8\sqrt{2}c_1^2c_2^2}\alpha_\varepsilon(\lambda').$$

Taking $c = 8\sqrt{2}c_1^2c_2^2$ finishes the proof. ■

Corollary 65 *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 21 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

$$\lambda_i = \begin{cases} \ell, & i \leq k, \\ e^{-\alpha(i-k)}, & i > k. \end{cases}, \quad \mu_i = \begin{cases} m_{0:k}, & i \leq k, \\ m_{k:\infty}e^{-\frac{\beta}{2}(i-k)}, & i > k. \end{cases}$$

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 21. 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 3, for Gaussian data the statement $\mathbb{P}(\mathscr{A}_k(L,\lambda)) \geq 1-\delta$ follows from the statement

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

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 21. For our covariance and regularization this translates into

$$\frac{1}{c_1}\frac{e^{-\alpha}}{1-e^{-\alpha}} \geq b\left(ne^{-\alpha} \vee \sqrt{n\frac{e^{-2\alpha}}{1-e^{-2\alpha}}}\right),$$

which can be equivalently rewritten as

$$1-e^{-\alpha} \leq \frac{1}{bc_1\sqrt{n}}\left(\frac{1}{\sqrt{n}} \wedge \sqrt{1-e^{-2\alpha}}\right).$$

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:

$$\alpha \leq \frac{1}{bc_1\sqrt{n}}\left(\frac{1}{\sqrt{n}} \wedge \sqrt{\alpha}\right),$$

that is

$$\alpha \leq \frac{1}{bc_1 n} \wedge \frac{1}{b^2c_1^2 n} = \frac{1}{b^2c_1^2 n}.$$

Let's take $c_2 = b^2c_1^2$ and put $\alpha = c_2^{-1}n^{-1}$. Then conditions from step 1 of Lemma 21 are satisfied.

The second part of Lemma 21 states that we require $n\lambda_k \geq C\Lambda(\lambda)$, that is, $n\ell \geq C\left(\frac{e^{-\alpha}}{1-e^{-\alpha}} + \lambda\right) = \frac{e^{-\alpha}}{1-e^{-\alpha}}$. Note that we also need $\ell \geq 1$ in order for the sequence $\{\lambda_i\}$ to be non-increasing. Given our previous choice of α we have

$$\frac{e^{-\alpha}}{1-e^{-\alpha}} \leq \frac{2}{\alpha} = 2c_2 n.$$

Thus, we can take $\ell = 2c_2$.

The third part of Lemma 21 requires the following

$$\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \leq C^{-1}n^{-1}\Lambda(\lambda)\|\boldsymbol{\mu}_{k:\infty}\|^2,$$

which we equivalently transform below:

$$\sum_{i>k} e^{-(\alpha+\beta)(i-k)} \leq C^{-2}n^{-1}\frac{e^{-\alpha}}{1-e^{-\alpha}}\sum_{i>k} e^{-\beta(i-k)},$$
$$\frac{e^{-(\alpha+\beta)}}{1-e^{-(\alpha+\beta)}} \leq C^{-2}n^{-1}\frac{e^{-\alpha}}{1-e^{-\alpha}}\frac{e^{-\beta}}{1-e^{-\beta}},$$
$$\frac{1-e^{-\beta}}{1-e^{-(\alpha+\beta)}} \leq C^{-2}n^{-1}\frac{1}{1-e^{-\alpha}}.$$

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:

$$\frac{2\beta}{\alpha+\beta}\le\frac{1}{C^2n\alpha}.$$

Plugging the expression for α yields

$$2\beta\le\frac{c_2}{C^2}\left(\beta+\frac{1}{c_2n}\right).$$

Actually, since $c_2=b^2C^2>2C^2$, 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 21 requires

$$nC^{-1}\Lambda(\lambda)^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2\ge\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}\ge n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}},$$

that is,

$$n\frac{1-e^{-\alpha}}{e^{-\alpha}}\frac{e^{-\beta}}{1-e^{-\beta}}\ge\frac{km_{0:k}^2}{\ell m_{k:\infty}^2}\ge n^2C^2\frac{(1-e^{-\alpha})^2}{e^{-2\alpha}}\frac{e^{-(\alpha+\beta)}}{1-e^{-(\alpha+\beta)}}.$$

Plugging in $\beta=\ln(2)$ and simplifying yields

$$n\frac{1-e^{-\alpha}}{e^{-\alpha}}\ge\frac{km_{0:k}^2}{\ell m_{k:\infty}^2}\ge n^2C^2\frac{(1-e^{-\alpha})^2}{e^{-\alpha}}\frac{1}{2-e^{-\alpha}}.$$

As before, let's replace it by a stronger condition:

$$\frac{n\alpha}{2}\ge\frac{km_{0:k}^2}{\ell m_{k:\infty}^2}\ge2n^2C^2\alpha^2.$$

Plugging in $\alpha=(c_2n)^{-1}$ and $\ell=2c_2$ yields

$$\frac{1}{2c_2}\ge\frac{km_{0:k}^2}{2c_2m_{k:\infty}^2}\ge\frac{2C^2}{c_2^2},$$
$$1\ge\frac{km_{0:k}^2}{m_{k:\infty}^2}\ge\frac{4C^2}{c_2}=\frac{4}{b^2}.$$

We see that since $b>4$, we can simply put $km_{0:k}^2/m_{k:\infty}^2=1$, and part 4 of Lemma 21 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 $n\Lambda^{-1}(\lambda)M(\lambda)$:

$$\begin{aligned}
\Lambda(\lambda) =& C^{-1}\frac{e^{-\alpha}}{1-e^{-\alpha}} \in \left(\frac{1}{2C\alpha}, \frac{2}{C\alpha}\right),\\
V(\lambda) :=& n^{-1}\mathrm{tr}\left(\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2}n\sum_{i>k}\lambda_i^2\\
=& n^{-1}\frac{k}{\left(1+n^{-1}\ell^{-1}C^{-1}\frac{e^{-\alpha}}{1-e^{-\alpha}}\right)^2} + C^2 n e^{2\alpha}(1-e^{-\alpha})^2\frac{e^{-2\alpha}}{1-e^{-2\alpha}}\\
\le& \frac{k/n}{\left(1+n^{-1}\ell^{-1}C^{-1}/(2\alpha)\right)^2} + C^2 n\frac{\alpha^2}{\alpha}\\
=& \frac{k/n}{\left(1+n^{-1}(2c_2)^{-1}C^{-1}c_2 n/2\right)^2} + C^2 n\alpha\\
=& \frac{k/n}{\left(1+0.25C^{-1}\right)^2} + C^2 c_2^{-1}\\
\le& 2.
\end{aligned}$$

$$\begin{aligned}
n\Lambda^{-1}(\lambda)M(\lambda) :=& \left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda(\lambda)^{-1}(\lambda)\|\boldsymbol{\mu}_{k:\infty}\|^2\\
=& k\frac{m_{0:k}^2}{\ell(1+n^{-1}\ell^{-1}\Lambda(\lambda))} + m_{k:\infty}^2 n\Lambda(\lambda)^{-1}\frac{e^{-\beta}}{1-e^{-\beta}}\\
\ge& k\frac{m_{0:k}^2}{\ell(1+2n^{-1}\ell^{-1}/(C\alpha))} + \frac{1}{2}m_{k:\infty}^2 nC\alpha(\lambda)\frac{e^{-\beta}}{1-e^{-\beta}}\\
=& \frac{km_{0:k}^2}{2c_2+2c_2/C} + \frac{1}{2}m_{k:\infty}^2 nC/(c_2 n)\\
\ge& \frac{1}{4c_2}(km_{0:k}^2 + m_{k:\infty}^2) = \frac{m_{k:\infty}^2}{2c_2}.
\end{aligned}$$

$$\begin{aligned}
n\Diamond^2(\lambda) :=& \left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2\\
=& k\frac{m_{0:k}^2}{\ell(1+n^{-1}\ell^{-1}\Lambda(\lambda))^2} + m_{k:\infty}^2 n^2\Lambda(\lambda)^{-2}\frac{e^{-(\alpha+\beta)}}{1-e^{-(\alpha+\beta)}}\\
=& \frac{2c_2 m_{k:\infty}^2}{(2c_2+\Lambda(\lambda)/n)^2} + m_{k:\infty}^2 n^2\Lambda(\lambda)^{-2}\frac{e^{-\alpha}}{2-e^{-\alpha}}
\end{aligned}$$

For $n\Diamond^2$ we need bounds from both sides. In what follows we write them separately.

$$\begin{aligned}
n\Diamond^2(\lambda) \le& \frac{2c_2 m_{k:\infty}^2}{(2c_2+0.5n^{-1}C^{-1}\alpha^{-1})^2} + m_{k:\infty}^2 n^2(2C\alpha)^2\\
=& \frac{2c_2 m_{k:\infty}^2}{(2c_2+0.5c_2C^{-1})^2} + m_{k:\infty}^2(2C/c_2)^2 \le \qquad 5m_{k:\infty}^2/c_2,
\end{aligned}$$

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

$$\begin{aligned} n\lozenge^2(\lambda) \geq & \frac{2c_2 m_{k:\infty}^2}{(2c_2 + 2n^{-1}C^{-1}\alpha^{-1})^2} + m_{k:\infty}^2 n^2 (2C\alpha)^2/3 \\ = & \frac{2c_2 m_{k:\infty}^2}{(2c_2 + 2c_2 C^{-1})^2} + m_{k:\infty}^2 (2C/c_2)^2/3 \\ \geq & m_{k:\infty}^2/(8c_2), \end{aligned}$$

where we used $2c_2 + 2c_2C^{-1} \leq 4c_2$ in the last transition.

Finally, we can write out the conditions from part 5 of Lemma 21. According to the bounds above, the following conditions on $m_{k:\infty}$ are sufficient:

$$m_{k:\infty}^2/(8c_2) \geq 2, \quad \frac{m_{k:\infty}^2}{2c_2} \geq a\sqrt{\frac{5}{c_2 n}} m_{k:\infty},$$

that is $m_{k:\infty}^2 \geq (16c_2) \vee (5c_2 a^2/n) = 16c_2$ given that n is large enough. ■

Lemma 66 *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 $\mathbb{P}(\mathscr{A}_k(L,\lambda)) \geq 1 - \delta$ 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*

$$\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}} \|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}} \geq C\|\boldsymbol{\mu}_{0:k}\|^2. \tag{30}$$

4. *Scale $\boldsymbol{\mu}$ up if needed, so that*

$$n\lozenge^2(\lambda) \geq V(\lambda) \quad \text{and } N(\lambda) \geq a\lozenge(\lambda).$$

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 11. 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 11 implies that

$$\alpha_\varepsilon(\lambda) \geq c_1^{-1} \frac{n\Lambda(\lambda)^{-1} M(\lambda)}{\sqrt{V(\lambda)} + \sqrt{n}\diamondsuit(\lambda)} \geq c_1^{-1} \frac{n\Lambda(\lambda)^{-1} M(\lambda)}{2\sqrt{n}\diamondsuit(\lambda)}.$$

At the same time, by Theorem 11 for any $\lambda' > \lambda$

$$\alpha_\varepsilon(\lambda') \leq c_1 \frac{n\Lambda(\lambda')^{-1} M(\lambda')}{\sqrt{V(\lambda')} + \sqrt{n}\diamondsuit(\lambda')} \leq c_1 \frac{n\Lambda(\lambda')^{-1} M(\lambda')}{\sqrt{n}\diamondsuit(\lambda')}.$$

Since $\|\boldsymbol{\mu}_{k:\infty}\| = 0$, $\Lambda(\lambda) \leq n\lambda_k$ and $\Lambda(\lambda') > n\lambda_1$, we can write

$$\begin{aligned}
n\Lambda(\lambda)^{-1} M(\lambda) =& \left\| \left(\Lambda(\lambda) n^{-1} \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\
\geq& \frac{1}{2} \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}, \\
n\Lambda(\lambda')^{-1} M(\lambda') =& \left\| \left(\Lambda(\lambda') n^{-1} \boldsymbol{I}_k + \boldsymbol{\Sigma}_{0:k}\right)^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\
\leq& 2n\Lambda(\lambda')^{-1} \|\boldsymbol{\mu}_{0:k}\|^2, \\
n\diamondsuit(\lambda)^2 \leq& \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}, \\
n\diamondsuit(\lambda')^2 =& \left\| \left(\Lambda(\lambda') n^{-1} \boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1} \boldsymbol{\Sigma}_{0:k}^{-1/2} \boldsymbol{\mu}_{0:k} \right\|^2 \\
\geq& \frac{1}{4} n^2 \Lambda(\lambda')^{-2} \|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}},
\end{aligned}$$

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

$$\begin{aligned}
\alpha_\varepsilon(\lambda) \geq c_1^{-1} \frac{n\Lambda(\lambda)^{-1} M(\lambda)}{\sqrt{n}\diamondsuit(\lambda)} \geq \frac{1}{2c_1} \frac{\frac{1}{2}\|\boldsymbol{\mu}_{0:k}\|^2_{\boldsymbol{\Sigma}_{0:k}^{-1}}}{\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}}} = \frac{1}{4c_1} \|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}} \geq \frac{C}{4c_1} \frac{\|\boldsymbol{\mu}_{0:k}\|^2}{\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}} = \\
= \frac{C}{16c_1} \frac{2n\Lambda(\lambda')^{-1} \|\boldsymbol{\mu}_{0:k}\|^2}{\frac{1}{2} n\Lambda(\lambda')^{-1} \|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}} \geq \frac{C}{16c_1^2} \cdot c_1 \frac{n\Lambda(\lambda')^{-1} M(\lambda')}{\sqrt{n}\diamondsuit(\lambda')} \geq \frac{C}{16c_1^2} \alpha_\varepsilon(\lambda').
\end{aligned}$$

Taking $c = 16c_1^2$ finishes the proof. ■

Corollary 67 *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 23 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:

$$\lambda_i = \begin{cases} e^{-\alpha i}, & i \le k, \\ \ell, & i > k. \end{cases}$$

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 23. When it comes to $\boldsymbol{\mu}_{0:k}$, we just put all its components to be equal, that is

$$\mu_i = \begin{cases} m, & i \le k, \\ 0, & i > k. \end{cases}$$

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 23 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

$$\ell(p-k)/c_1 = \Lambda(\lambda) \ge c_1\left(n\lambda_{k+1} \vee \sqrt{n\sum_{i>k}\lambda_i^2}\right) = c_1\ell(n \vee \sqrt{n(p-k)}).$$

Due to Lemma 3, 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 23 requires $n\lambda_i > \Lambda(\lambda)$, that is, $ne^{-\alpha k} \ge \ell(p-k)/c_1$, so we can take $\ell = c_1 ne^{-\alpha k}/p$. 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 23 demands $\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}}\|\boldsymbol{\mu}_{0:k}\|_{\boldsymbol{\Sigma}_{0:k}^{-1}} \ge C\|\boldsymbol{\mu}_{0:k}\|^2$, that is,

$$\sqrt{\left(m^2\sum_{i=1}^k e^{-\alpha i}\right)\left(m^2\sum_{i=1}^k e^{\alpha i}\right)} \ge c_1 k m^2,$$

$$\sqrt{\frac{1-e^{-k\alpha}}{1-e^{-\alpha}}\frac{e^{k\alpha}-1}{e^{\alpha}-1}} \ge c_1 k.$$

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:

$$e^{(k-1)\alpha/2} \ge c_1 k,$$

so we need to take $\alpha \ge 2\ln(c_1 k)/(k-1)$. 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 $\ell = c_1 ne^{-\alpha k}/(ep) = c_1 n/(epk^4)$. 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 23 is satisfied. To check that part, we start with writing the expressions for $n\Diamond^2(\lambda), V(\lambda)$, 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 $\Diamond$, the derivation is very similar as above:

$$\begin{aligned}
n\Diamond^2(\lambda) :=& \left\|\left(\Lambda(\lambda)n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1}+\boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2+n^2\Lambda(\lambda)^{-2}\|\boldsymbol{\mu}_{k:\infty}\|^2_{\boldsymbol{\Sigma}_{k:\infty}}\\
=&m^2\sum_{i=1}^{k}\frac{e^{\alpha i}}{\left(1+e^{\alpha i}(p-k)/(epk^4)\right)^2}
\end{aligned}$$

For $n\Diamond^2$ we need bounds from both sides. In what follows we write them separately.

$$\begin{aligned} n\Diamond^2(\lambda) \leq & m^2 \sum_{i=1}^{k} \frac{e^{\alpha i}}{1} \\ \leq & \frac{m^2 k^5}{2\ln(k)}, \\ n\Diamond^2(\lambda) \geq & \frac{m^2}{\left(1+e^{\alpha k}/(ek^4)\right)^2} \sum_{i=1}^{k} e^{\alpha i} \\ = & \frac{m^2}{(1+e^{-1})^2} \sum_{i=1}^{k} e^{\alpha i} \\ \geq & \frac{m^2 k^5}{20\ln(k)}. \end{aligned}$$

Finally, to satisfy part 4 of Lemma 23 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

$$\frac{m^2 k^5}{20\ln(k)} \geq \frac{k}{n} + \frac{n}{p-k}, \quad \frac{m^2 k^5}{10\ln(k)} \geq a_1 \sqrt{\frac{m^2 k^5}{2n\ln(k)}},$$

which is equivalent to

$$\frac{m^2 k^5}{20\ln(k)} \geq \left(\frac{k}{n} + \frac{n}{p-k}\right) \vee \frac{50 a_1}{n}.$$

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. ■

Appendix L. Comparisons with earlier results

Proposition 68 *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

$$\begin{aligned}
\sum_i \lambda_i^2 \leq& \Lambda,\\
\Lambda =& \sum_i \lambda_i \in [\kappa p, p],\\
V =& \Lambda^{-2} n \sum_{i>k} \lambda_i^2 \leq n/\Lambda \leq \frac{n}{\kappa p},\\
\Delta V \leq& \frac{n\lambda_1^2}{\Lambda^2} + \frac{n\lambda_1^2 + \sum_i \lambda_i^2}{\Lambda^2} \leq \frac{2n}{\kappa^2 p^2} + \frac{1}{\kappa p} \leq \frac{3}{\kappa^2 p},\\
\lozenge^2 =& n\Lambda^{-2}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2 \leq \frac{n\|\boldsymbol{\mu}\|^2}{\kappa^2 p^2},\\
M =& \|\boldsymbol{\mu}\|^2.
\end{aligned}$$

Plugging in those bounds together with $\sigma_\eta < 1$ yields

$$\begin{aligned}
n\Lambda^{-1} M - ct\lozenge \geq& \frac{n}{p}\|\boldsymbol{\mu}\|^2 - ct\frac{\sqrt{n}\|\boldsymbol{\mu}\|}{\kappa p},\\
\left[1 + n\Lambda^{-1} M\sigma_\eta\right]\sqrt{V + t^2\Delta V} + \lozenge\sqrt{n} \leq& \left[1 + \frac{n}{\kappa p}\|\boldsymbol{\mu}\|^2\right]\sqrt{\frac{n}{\kappa p} + t^2\frac{3}{\kappa^2 p}} + \frac{n\|\boldsymbol{\mu}\|}{\kappa p}.
\end{aligned}$$

Next, since $n\|\boldsymbol{\mu}\|^2/p \leq \kappa$ for $t^2 \leq n\kappa$ we can write

$$\left[1 + n\Lambda^{-1} M\sigma_\eta\right]\sqrt{V + t^2\Delta V} + \lozenge\sqrt{n} \leq [1+1]\sqrt{\frac{4n}{\kappa p}} + \frac{n\|\boldsymbol{\mu}\|}{\kappa p} \leq 5\sqrt{\frac{n}{p\kappa}}.$$

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 69 *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 $t < \sqrt{n}$,*

$$\frac{N - ct\lozenge}{\sqrt{V + t^2\Delta V} + \sqrt{n}\lozenge} \geq \frac{1}{4}\frac{n\|\boldsymbol{\mu}\|^2}{n\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} + \sqrt{n}\|\boldsymbol{\Sigma}\|_F + n\|\boldsymbol{\Sigma}\|}. \tag{34}$$

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 $t \le \sqrt{n}$ the condition $\|\boldsymbol{\mu}\|^2 \ge 2c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$ 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 70 *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 29. Then for $t \le \sqrt{n}$*

$$\frac{N - ct\Diamond}{\sqrt{V + t^2\Delta V} + \sqrt{n}\Diamond} \ge \frac{1}{6}\frac{\|\boldsymbol{\mu}\|^2}{A + B + \lambda_j + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}.$$

Proof Note that under assumption $\|\boldsymbol{\mu}_{0:k}\| = 0$ we have

$$M = \|\boldsymbol{\mu}_{k:\infty}\|^2 = \|\boldsymbol{\mu}\|^2, \quad \Diamond^2 = n\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 = n\Lambda^{-2}\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2,$$

and thus, the condition $\|\boldsymbol{\mu}\|^2 \ge 2c\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}$ can be rewritten as $M \ge 2c\sqrt{n}\Lambda\Diamond$. Therefore, it implies that $n\Lambda^{-1}M - ct\Diamond \ge n\Lambda^{-1}M/2$ for $t \le \sqrt{n}$. Thus,

$$\frac{n\Lambda^{-1}M - ct\Diamond}{\sqrt{V + t^2\Delta V} + \sqrt{n}\Diamond} \ge \frac{1}{2}\frac{\|\boldsymbol{\mu}\|^2}{n^{-1}\Lambda\sqrt{V + t^2\Delta V} + \|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}}. \tag{56}$$

We see that in order to compare Equation (56) and (38), 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 71 *Take real q,r,s such that $0\leq r<1<s$, $0\leq q<s-r$. Consider $p=n^s$, $\boldsymbol{\Sigma}=\mathrm{diag}(\lambda_1,\ldots,\lambda_p)$, and $\boldsymbol{\mu}=\sqrt{2\lambda_1/\pi}\boldsymbol{e}_1$, where $\{\lambda_i\}_{i=1}^p$ are given by Equation (40). 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.

$$\begin{aligned}
\boldsymbol{\mu} :=& \sqrt{\frac{2\lambda_1}{\pi}}\boldsymbol{e}_1 = \sqrt{2/\pi}n^{(s-q-r)/2}\boldsymbol{e}_1, \\
\Lambda =& \sum_{i>k}\lambda_i = (n^s - n^r)\lambda_{k+1} = (n^s - n^r)\cdot(1-n^{-q})/(1-n^{r-s}) \\
=& n^s - n^{s-q} = n^s(1-o(1)), \\
\Lambda n^{-1}\lambda_1^{-1} =& n^s(1-o(1))\cdot n^{-1}\cdot n^{-s+q+r} = n^{q+r-1}(1-o(1)), \\
\diamondsuit^2 =& n^{-1}\left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda^{-2}\|\boldsymbol{\mu}_{k:\infty}\|_{\boldsymbol{\Sigma}_{k:\infty}}^2 \\
=& n^{-1}\cdot(1+\Lambda n^{-1}\lambda_1^{-1})^{-2}\lambda_1^{-1}\cdot\frac{2\lambda_1}{\pi} \\
=& \frac{2+o(1)}{\pi n\left(1+n^{q+r-1}\right)^2}, \\
n\Lambda^{-1}M =& \left\|\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-1/2}\boldsymbol{\Sigma}_{0:k}^{-1/2}\boldsymbol{\mu}_{0:k}\right\|^2 + n\Lambda^{-1}\|\boldsymbol{\mu}_{k:\infty}\|^2 \\
=& (1+\Lambda n^{-1}\lambda_1^{-1})^{-1}\lambda_1^{-1}\cdot\frac{2\lambda_1}{\pi} \\
=& \frac{2+o(1)}{\pi(1+n^{q+r-1})}, \\
V =& n^{-1}\mathrm{tr}\left(\left(\Lambda n^{-1}\boldsymbol{\Sigma}_{0:k}^{-1} + \boldsymbol{I}_k\right)^{-2}\right) + \Lambda^{-2}n\sum_{i>k}\lambda_i^2 \\
=& n^{-1}n^r(\Lambda n^{-1}\lambda_1^{-1}+1)^{-2} + \Lambda^{-2}n(n^s-n^r)\lambda_{k+1}^2 \\
=& \frac{1+o(1)}{n^{1-r}(1+n^{q+r-1})^2} + n^{1-s}(1+o(1)), \\
\Delta V =& \frac{1}{n}\wedge\frac{n\lambda_1^2}{\Lambda^2} + \frac{n\lambda_{k+1}^2 + \sum_{i>k}\lambda_i^2}{\Lambda^2} \\
=& n^{-1}\wedge n^{1-2r-2q}(1+o(1)) + n^{-s}(1+o(1)).
\end{aligned}$$

Now let's plug this into the quantity of interest. Note that as long as $t = o(\sqrt{n})$, we have $t\diamondsuit = o(n\Lambda^{-1}M)$. 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 $n\Lambda^{-1}M = \diamondsuit\sqrt{n}(\sqrt{2/\pi}+o(1))$. That is,

$$\frac{n\Lambda^{-1}M}{\sqrt{V}+\sqrt{n}\diamondsuit} = \frac{\sqrt{2/\pi}+o(1)}{1+\sqrt{V/(n\diamondsuit^2)}}.$$

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. ■

References

P. Abdalla and N. Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *Preprint, arXiv:2205.08494*, 2023.

N. Ardeshir, C. Sanford, and D. J. Hsu. Support vector machines and linear regression coincide with very high-dimensional features. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4907–4918. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/26d4b4313a7e5828856bc0791fca39a2-Paper.pdf`.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

M. Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Y. Cao, Q. Gu, and M. Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *CoRR*, abs/2104.13628, 2021. URL `https://arxiv.org/abs/2104.13628`.

N. S. Chatterji and P. M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

S. Frei, G. Vardi, P. L. Bartlett, and N. Srebro. Benign overfitting in linear classifiers and leaky ReLU networks from KKT conditions for margin maximization. In *Conference on Learning Theory*, 2023.

D. Hsu, V. Muthukumar, and J. Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 91–99, 2021.

H. Huang and Q. Yang. Large dimensional analysis of general margin based classification methods. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(11):113401, nov 2021. doi: 10.1088/1742-5468/ac2edd. URL `https://dx.doi.org/10.1088/1742-5468/ac2edd`.

B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

A. Montanari, F. Ruan, Y. Sohn, and J. Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *Preprint, arXiv:1911.01544*, 2023.

V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

O. Shamir. The implicit bias of benign overfitting. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 448–478. PMLR, 02–05 Jul 2022. URL `https://proceedings.mlr.press/v178/shamir22a.html`.

K. Tikhomirov. Sample covariance matrices of heavy-tailed distributions. *International Mathematics Research Notices*, 2018(20):6254–6289, 2018.

A. Tsigler. *Benign Overfitting in Linear Regression and Classification*. PhD thesis, University of California, Berkeley, 2024. URL `https://escholarship.org/uc/item/6z11k9j3`.

A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023. URL `http://jmlr.org/papers/v24/22-1398.html`.

R. 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.

K. Wang and C. Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting and regularization. *Preprint, arXiv:2011.09148*, 2021.

K. Wang, V. Muthukumar, and C. Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=0O5jpovbdH0`.