Minimax Analysis of Estimation Problems in Coherent Imaging

Hao Xing†, Soham Jana†, and Arian Maleki

Abstract

Unlike conventional imaging modalities, such as magnetic resonance imaging, which are often well described by a linear regression framework, coherent imaging systems follow a significantly more complex model. In these systems, the task is to estimate the unknown image $\boldsymbol{x}_o \in \mathbb{R}^n$ from observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L \in \mathbb{R}^m$ of the form

$$\boldsymbol{y}_l = A_l X_o \boldsymbol{w}_l + \boldsymbol{z}_l, \quad l = 1, \ldots, L,$$

where $X_o = \text{diag}(\boldsymbol{x}_o)$ is an $n \times n$ diagonal matrix, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_L \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_n)$ represent speckle noise, and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_z^2 I_m)$ denote additive noise. The matrices $A_1, \ldots, A_L$ are known forward operators determined by the imaging system.

The fundamental limits of conventional imaging systems have been extensively studied through sparse linear regression models. However, the limits of coherent imaging systems remain largely unexplored. Our goal is to close this gap by characterizing the *minimax risk of estimating* $\boldsymbol{x}_o$ in high-dimensional settings.

Motivated by insights from sparse regression, we observe that the *structure* of $\boldsymbol{x}_o$ plays a crucial role in determining the estimation error. In this work, we adopt a general notion of structure based on the covering numbers, which is more appropriate for coherent imaging systems. We show that the minimax mean squared error (MSE) scales as

$$\frac{\max\{\sigma_z^4,\, m^2,\, n^2\}\, k \log n}{m^2 n L},$$

where k is a parameter that quantifies the effective complexity of the class of images.

1 Introduction

1.1 Motivation and main objective

Coherent imaging technology, which uses coherent light sources such as lasers to illuminate the object of interest, underpins many modern imaging systems. Examples include Optical Coherence Tomography (OCT) Schmitt et al. (1999), ultrasound imaging Achim et al. (2001), Synthetic Aperture Radar (SAR) Lopez-Martinez and Fabregas (2003); Dasari et al. (2015), digital holography Bianco et al. (2018), and near-infrared spectroscopy (NIRS) Ortega-Martinez et al. (2019). Compared to other imaging modalities, coherent imaging systems are affected by a complex form of distortion known as speckle noise Racine et al. (1999).

†Equal contributions. H. X. is with the Department of Mathematics, City University of New York - the Graduate Center, NY, USA (email:hao.xing43@gc.cuny.edu). S. J. is with the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, IN, USA, (email:soham.jana@nd.edu). A. M. is with the Department of Statistics, Columbia University, NY, USA.(email: mm4338@columbia.edu)

The general mathematical problem that arises in coherent imaging systems is to estimate a signal or image $\boldsymbol{x}_o \in \mathbb{R}_+^n$ from measurements $\boldsymbol{y}$ of the form

$$\boldsymbol{y} = AX_o\boldsymbol{w} + \boldsymbol{z}, \tag{1.1}$$

where $X_o \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose diagonal elements are the same as $\boldsymbol{x}_o$, $\boldsymbol{w} \in \mathbb{R}^n$, and $\boldsymbol{z} \in \mathbb{R}^m$ represent the speckle and additive noises, respectively. In this model, $A \in \mathbb{R}^{m\times n}$ is a known matrix, called the forward operator of the imaging system. Note that, compared to linear regression problems, which are popular for other types of imaging such as MRI and CT, the relationship between $\boldsymbol{y}$ and $\boldsymbol{x}_o$ is further distorted by the speckle noise $\boldsymbol{w}$. In many applications, the speckle noise is "fully developed," which means that the elements of $\boldsymbol{w}$ are i.i.d. $N(0, 1)$. A standard model for the additive noise $\boldsymbol{z}$ is also that it has i.i.d. $N(0, \sigma_z^2)$.

Before discussing our modeling assumptions, we first highlight an essential technique in coherent imaging systems, namely *multilook* or *multishot* measurements. It is widely recognized in the coherent imaging community that estimating $\boldsymbol{x}_o$ from a single measurement of the form in (1.1) is challenging, and in most applications the reconstruction quality is insufficient. Therefore, in many settings, such as SAR and digital holography De Vries (1998); Argenti et al. (2013); Bate et al. (2022), multiple measurements of the same scene are acquired. More specifically, one collects measurements $\boldsymbol{y}_1, \dots, \boldsymbol{y}_L$ of the form

$$\boldsymbol{y}_l = A_l X_o \boldsymbol{w}_l + \boldsymbol{z}_l, \quad l = 1, 2, .., L \tag{1.2}$$

where L is referred to as the number of *looks*, and $A_1, \dots, A_L$ represent the forward operators of different shots. There are a few points that we should clarify about this multilook system:

- In practice, effort is made to ensure that $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_L$ are independent of each other. Similarly, the additive noise vectors $\boldsymbol{z}_1, \dots, \boldsymbol{z}_L$ are typically assumed to be independent across measurements and also independent of $\boldsymbol{w}_1, \dots, \boldsymbol{w}_L$.

- The forward models across looks may differ or be the same, depending on the technology used. For instance, if the wavelength of the illuminating light changes, then different A_i's will be observed, but if phase masks are used on the path of the illuminating light, then we will have $A_1 = A_2 = \dots = A_L$.

In our mathematical model, we assume that the variance of the speckle noise is equal to 1. This is without any loss of generality. Consider the case where the standard deviations of the multiplicative noise and the additive noise are σ_w and σ_z, respectively. Then, by dividing the sensor measurements $\boldsymbol{y}_l$ by σ_w, we obtain an equivalent system of measurements:

$$\tilde{\boldsymbol{y}}_l = AX_o\tilde{\boldsymbol{w}}_l + \tilde{\boldsymbol{z}}_l.$$

Here we have defined $\tilde{\boldsymbol{y}}_l := \frac{\boldsymbol{y}_l}{\sigma_w}, \tilde{\boldsymbol{w}}_l := \frac{\boldsymbol{w}_l}{\sigma_w}, \tilde{\boldsymbol{z}}_l := \frac{\boldsymbol{z}_l}{\sigma_w}$. As a result of this transformation, we have

$$\tilde{\boldsymbol{w}}_l \sim \mathcal{N}(0, I), \quad \tilde{\boldsymbol{z}}_l \sim \mathcal{N}\left(0, \frac{\sigma_z^2}{\sigma_w^2} I\right),$$

which is consistent with (1.2). Hence, without a loss of generality, we set the variance of the speckle noise to 1 and discuss the changes in σ_z.

Despite the widespread use of coherent imaging technology across many applications, the theoretical aspects of the associated estimation problems remain largely unexplored. The main goal of this paper is to help fill this gap by addressing the following questions:

1. How do m, n, L and σ_z^2 affect the accuracy of the estimates of $\boldsymbol{x}_o$?
2. Is there any gain in using different $A_1, A_2, \ldots, A_L$ compared to the case $A_1 = A_2 = \ldots = A_L$?
3. How do additive and multiplicative noise compare in their impact on the estimation error?

To address these questions and to provide insight into the estimation challenges that arise in coherent imaging systems, we seek to characterize the minimax risk associated with this problem. In particular, we aim to quantify

$$R_2(\mathcal{C}, m, n, \sigma_z^2) := \inf_{\widehat{\boldsymbol{x}}} \sup_{\boldsymbol{x}_o \in \mathcal{C}} \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2}{n}\right], \tag{1.3}$$

where $\widehat{\boldsymbol{x}}$ is any measurable estimate that has access to $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L$ and $A_1, A_2, \ldots, A_L$, and $\mathcal{C}$ denotes the set of all possible options for $\boldsymbol{x}_o$. Prior work in sparse linear regression and compressed sensing shows that the minimax risk is strongly influenced by the choice of $\mathcal{C}$. In the next section, we first describe our choice of $\mathcal{C}$, and then discuss our contributions and our responses to the questions we raised above.

1.2 Notations

Throughout the paper, for the sake of clarity, all matrices of sizes $m \times n$, $m \times m$, and $n \times n$ (without dependence on the number of looks L) are represented by uppercase italic letters such as A_l, Σ_l, and X. We use boldface uppercase letters, e.g. $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{\Sigma}$, when the sizes of matrices depend on L. These matrices are often constructed by stacking matrices of smaller sizes and may have sizes such as $mL \times (m+n)L$ and $(m+n)L \times (m+n)L$. For a matrix A, $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ denote the maximum and minimum singular values of A. Furthermore $\|A\|_2 = \sigma_{\max}(A)$ and $\|A\|_{\text{HS}}$ denote the spectral norm and Hilbert-Schmidt norms of A, respectively. Boldface lowercase letters such as $\boldsymbol{x}$ are used for vectors for sizes m or n (again, no dependence on L). Arrows above the vectors emphasize that the dimensions of the vectors depend on L. Again, such vectors are constructed by stacking L lower-dimensional vectors. For a vector $\boldsymbol{x} = (x_1, ..., x_n)$, we let $\boldsymbol{x}^2 := (x_1^2, ..., x_n^2)$. Given two sequences $\{a_n\}_{n \in \mathbb{N}}$ and $\{b_n\}_{n \in \mathbb{N}}$, we say $a_n = O(b_n)$, or equivalently $a_n \ll b_n$, $b_n = \Omega(a_n)$, if there exist constants $C > 0$, and $M > 0$, such that for all $n > M$, $|a_n| \leq C|b_n|$. We say $a_n = o(b_n)$, or equivalently $b_n = \omega(a_n)$, if $\lim_{n \to \infty} a_n/b_n = 0$. We write $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. For $a, b \in \mathbb{R}$, we denote $(a, b]_{\mathbb{Z}} := \{x \in Z : a < x \leq b\}$.

1.3 Organization of the paper

In Section 2, we introduce our model, state the main assumptions, and present our primary contributions. Section 3 reviews related work and compares our results with the existing literature. In Section 4, we provide the necessary preliminaries for our analysis. The remaining sections are devoted to the proofs of the theorems stated in Section 2.

2 Our main contributions

In this section, we first discuss our choice of the set $\mathcal{C}$ in (1.3), and then present our theoretical results in response to the questions we raised in Section 2.2.

2.1 The choice of $\mathcal{C}$

Inspired by developments in the fields of sparse regression and compressed sensing, we note that the structure of $\boldsymbol{x}_o$, plays a crucial role in determining the accuracy of the estimates. As will be clarified later in this section, sparsity is not useful for coherent imaging systems. Hence, in this paper we work with a more general notion of "structuredness". This notion allows us to cover not only the class of k-sparse vectors, but also the more modern classes developed in the field of neural networks, such as the class of untrained networks. For a compact set $\mathcal{C} \subset \mathbb{R}^n$, let $N_\varepsilon(\mathcal{C})$ denote its covering number under the ℓ_2 metric, namely the least number of ℓ_2-balls covering $\mathcal{C}$..

Definition 2.1. We say that $\mathcal{C} \subset \mathbb{R}^n$ satisfies polynomial complexity of order k if there exist constants $a > 0, b \geq 0$ independent of k and n such that

$$N_\varepsilon(\mathcal{C}) \leq \left(\frac{an^b}{\varepsilon}\right)^k. \tag{2.1}$$

Before proceeding, we review several sets with polynomial complexity of order k to establish the usefulness of this definition. The proof of the following results are provided in Appendix A. Our first example can serve as a proxy for images generated by neural network architectures such as deep image priors Ulyanov et al. (2018), implicit neural representations Sitzmann et al. (2020), and autoencoders Bank et al. (2023). These models have been extensively used as reliable and accurate models for images.

Example 2.2. Consider $k \ll n$ and let $g : \mathbb{R}^k \to \mathbb{R}^n$ denote a Lipschitz function with a Lipschitz constant M. Define

$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} = g(\theta) \ \text{ for some } \theta \in [0,1]^k\}.$$

Then, $N_\varepsilon(\mathcal{C}) \leq \left(\frac{2M\sqrt{k}}{\epsilon} + 1\right)^k$. Note that when $\epsilon < 2M\sqrt{k}$, we have $\frac{2M\sqrt{k}}{\epsilon} + 1 < \frac{3M\sqrt{k}}{\epsilon}$, and hence we can also have $N_\varepsilon(\mathcal{C}) \leq \left(\frac{3M\sqrt{k}}{\epsilon}\right)^k$.

Definition 2.2 covers a wide range of examples. For instance, in the literature of neural networks, it has been conjectured that the output of certain neural networks, such as implicit neural representation networks Sitzmann et al. (2020) and deep image priors Heckel and Hand (2018) can generate all natural images as the parameters of the networks change. Note that the number of parameters of these networks can be interpreted as k in Example 2.2. Often times the number of parameters is much smaller than the ambient dimension of the signal that is generated by the network. Our second example offers upper bounds for the covering numbers of k-sparse vectors.

Example 2.3. For the set $C = B_2(1) \cap \mathcal{S}_k$, where $\mathcal{S}_k = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_0 \leq k\}$, we have

$$\left(\frac{1}{\epsilon}\right)^k \leq N_\epsilon(\mathcal{C}) \leq \binom{n}{k}\left(\frac{2}{\varepsilon} + 1\right)^k \leq \left(\frac{2n}{\varepsilon} + n\right)^k \leq \left(\frac{3n}{\varepsilon}\right)^k.$$

The following example is a slight generalization of the above example that can cover a wide range of models.

Example 2.4. Let $D \in \mathbb{R}^{n \times n}$ denote a matrix with the maximum singular value $\sigma_{\max}(D)$. Suppose that $\mathcal{C} \subset \{D\theta \mid \theta \in \mathcal{S}_k \cap B_2(0,1)\}$. Then, $N_\varepsilon(\mathcal{C}) \leq \left(\frac{3\sigma_{\max}(D)n}{\epsilon}\right)^k$.

Note that simple applications of the above result are piecewise constant and piecewise polynomial functions. The following lemma, proves this claim for the class of piecewise constant vectors.

Example 2.5. Define $D_{(1)} \in \mathbb{R}^{n\times n}$ as a matrix whose diagonal elements are equal to 1 and the immediate super diagonal elements are equal to -1. Suppose that for every $\boldsymbol{x} \in \mathcal{C}$, $D_{(1)}\boldsymbol{x} \in \mathcal{S}_k \cap B_2(0,1)$. Then,

$$N_\varepsilon(\mathcal{C}) \leq \left(\frac{3n^2}{\varepsilon}\right)^k.$$

Note that if $\boldsymbol{x}$ is a constant vector, meaning that all its entries have the same value, then all elements of $D_{(1)}\boldsymbol{x}$ except for the last one are equal to zero. Therefore, if we assume that $D_{(1)}\boldsymbol{x} \in \mathcal{S}_k$, it follows that $\boldsymbol{x}$ is a piecewise constant vector with at most k jumps (changes) in its values.

There are several ways to extend the above example to piecewise polynomial functions of degree at most P. We do the following simple extension.

Example 2.6. Define $D_{(1)} \in \mathbb{R}^{n\times n}$ as a matrix whose diagonal elements are equal to 1 and the immediate super diagonal elements are equal to -1. Define $D_{(P+1)} = (D_{(1)})^{P+1}$. Suppose that for every $\boldsymbol{x} \in \mathcal{C}$, $D_{(P+1)}\boldsymbol{x} \in \mathcal{S}_k \cap B_2(0,1)$. Then,

$$N_\varepsilon(\mathcal{C}) \leq \left(\frac{3n^{P+2}}{\varepsilon}\right)^k.$$

Note that if $f : [0,1] \to \mathbb{R}$ is a polynomial of degree P, and $\boldsymbol{x}_i = f(i/n)$, it follows from a well-known fact in the theory of forward difference operators (see (Graham et al., 1994, Section 5.3)) that all elements of $D_{(P+1)}(\boldsymbol{x})$ are equal to zero, except possibly for the last $P+1$ elements. Hence, the set $\mathcal{C}$ in Example 2.6 can be viewed as discretized piecewise polynomial vectors.

Inspired by all the examples above, in our theoretical results we will be assuming that $\boldsymbol{x}_o$ in (1.1) is from a set $\mathcal{C}$ that has a polynomial complexity of order $k \ll n$.

2.2 Main theoretical result for independent A_is

As we discussed before, we consider the problem of estimating $\boldsymbol{x}_o$ from the observations

$$\boldsymbol{y}_l = A_l X_o \boldsymbol{w}_l + \boldsymbol{z}_l, \text{ for } l = 1, \dots, L, \tag{2.2}$$

under the assumption $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_L \overset{i.i.d.}{\sim} N(0, I)$, and $\boldsymbol{z}_l \overset{i.i.d.}{\sim} N(0, \sigma_z^2 I)$. Our main goal is to characterize the minimax risk of the estimation problem in (2.2) defined as:

$$R_2(\mathcal{C}, m, n, \sigma_z) := \inf_{\widehat{\boldsymbol{x}}} \sup_{x_o \in \mathcal{C}} \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2}{n}\right]. \tag{2.3}$$

Our first theorem obtains an upper bound for this quantity: Let $\mathcal{F}_{a,b,k,n}$ denote all subsets of $[x_{\min}, x_{\max}]^n$ whose ε-covering number is upper bounded by $\left(\frac{an^b}{\varepsilon}\right)^k$.

Theorem 2.7. Suppose that $A_1, ..., A_L$ are independent $m \times n$ matrices and have i.i.d. $N(0,1)$ entries. Suppose that $\boldsymbol{x}_o \in \mathcal{C}_k \in \mathcal{F}_{a,b,k,n}$. If $mL \leq n^4 k \log n$, then

$$R_2(\mathcal{C}_k, m, n, \sigma_z) = O_{x_{\max}, x_{\min}, a, b}\left(\min\left(\frac{\max(\sigma_z^4, m^2, n^2)k \log n}{m^2 nL}, 1\right)\right). \tag{2.4}$$

The above theorem obtains an upper bound for the minimax risk that holds for any $\mathcal{C}_k \subset [x_{\min}, x_{\max}]^n$ that satisfies polynomial complexity of order k. Before discussing the assumptions made in the above theorem, let us discuss the sharpness of this upper bound.

Theorem 2.8. Suppose $a \geq x_{\max} - x_{\min}$ and $b \geq 1$. If $\log m = \Theta(\log n)$, $\log L = O(\log n)$, and there exists $\varepsilon \in (0, 1/2)$ such that $k \leq n^{1-2\varepsilon}$, and that $\max(\sigma_z^4, m^2, n^2) k \log n \leq m^2 n^{1-\varepsilon} L$, then we have

$$\sup_{\mathcal{C} \in \mathcal{F}_{a,b,k,n}} R_2(\mathcal{C}, m, n, \sigma_z) = \Omega_{\varepsilon, x_{\max}, x_{\min}} \left(\frac{\max(\sigma_z^4, m^2, n^2) k \log n}{m^2 n L} \right). \tag{2.5}$$

Remark 2.9. One can easily extend Definition 2.8 to any $a, b > 0$. We shall provide rationale in Section B.

Before we discuss the implications of our result, we discuss some of the assumptions we have made in the above theorems. A natural question is why we did not adopt the standard notion of sparsity widely used in sparse regression and imaging systems that fit well within the framework of linear regression. We mention two reasons below:

1. In imaging sciences, it is often the case that the vector $\boldsymbol{x}$ is not sparse itself. In fact, some linear transformation of the vector, e.g. wavelet or Fourier transform of $\boldsymbol{x}$ is sparse Donoho et al. (1995); Donoho and Johnstone (1998a). Suppose that $\boldsymbol{x} = F\boldsymbol{u}$, where $\|\boldsymbol{u}\|_0 \leq k$. Then, in linear regression, one can write the measurement $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{z}$ as $\boldsymbol{y} = \tilde{A}\boldsymbol{u} + \boldsymbol{z}$, where $\tilde{A} = AF$. Hence, the problem of imaging when linear model is accurate, is equivalent to the problem of sparse linear regression. As is clear, because of the nature of the speckle noise, we **cannot** transform the estimation of $\boldsymbol{x}$ from the observation $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_L$ to an estimation of a sparse vector (with a different design matrix).

2. Because of the nature of speckle noise (that is multiplied by $\boldsymbol{x}$), the estimation of sparse signals are easier than the estimation of the non-sparse signals. Intuitively speaking, this is due to the fact that sparse vectors, automatically remove most of the speckle noises. In other words, out of the n-speckle noise elements that are often present in these systems, $n - k$ of them will be multiplied by zeros during the measurement process and will not have a major impact on the estimation problem. To confirm this intuition rigorously, the next theorem shows that the minimax risk of estimating sparse vectors from (2.2) is much smaller than the bounds presented in Theorems 2.7 and 2.8.

Let

$$\mathcal{S}_k^{\text{bdd}} := \left\{ \boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_0 \leq k, x_i = 0 \text{ or } 0 < x_{\min} \leq x_i \leq x_{\max} \right\}. \tag{2.6}$$

Theorem 2.10. If $k \log(en/k) \leq m$, then there exist constants $c_{x_{\max}, x_{\min}}, C_{x_{\max}, x_{\min}}$ only depending on $x_{\max}$ and $x_{\min}$ such that

$$c_{x_{\max}, x_{\min}} \frac{k}{nL} \leq R_2(\mathcal{S}_k^{\text{bdd}}, m, n, k, L, \sigma_z) \leq C_{x_{\max}, x_{\min}} \left(\frac{k}{nL} + \frac{\sigma_z^2 k \log(n/k)}{mn} \right) \tag{2.7}$$

In particular, if $\sigma_z^2 L \log(n/k) \leq m$, the upper and lower bounds have the same order $\frac{k}{nL}$.

The proof of this result can be found in Section G. By comparing Theorem 2.10 with Theorems 2.7 and 2.8, it is straightforward to see that estimating k-sparse signals can be much easier than estimating other types of signals that satisfy polynomial growth of order k that includes for instance, the signals that are sparse in an orthonormal basis such as wavelet (see definition 2.4).

The remaining assumptions of Theorems 2.7 and 2.8 are only technical and relatively minor; they could likely be removed, though doing so would make the proof less transparent. We discuss these assumptions briefly below:

1. $mL \leq n^4 k \log n$: This assumption has appeared in Theorem 2.7. In practice, obtaining more than $L > 100$ independent looks is rarely feasible. Since in most imaging applications n is on the order of hundreds of thousands to millions, this condition is typically satisfied.

2. $\log(L) = O(\log n)$: This assumption appears in Definition 2.8. As we discussed before, in all applications, L is much smaller than n. Hence, the assumption that L is not growing too fast in terms of n is a natural assumption in practice.

3. $\log m = \Theta(\log n)$: As will be discussed in the proof of definition 2.8 and might be even clear from the formulation of the problem, increasing m beyond n does not help in removing the speckle noise, and it only helps in removing the additive part of the noise. Hence, increasing m to a very large number is not particularly helpful in reducing the risk, since unless the additive noise is too large, the errors induced by the speckle noise are the dominant part of the risk. Note that increasing m in imaging applications is equivalent to increasing the number of sensors which is costly. As a result, there is no reason to increase m much beyond n in real-world applications, and again the assumption $\log m = \Theta(\log n)$ is a mild assumption.

4. $k \leq n^{1-2\varepsilon}$: This assumption is used in definition 2.8. It is always the case that $k \ll n$. Hence, again this is a mild assumption. However, at this stage it is unclear, whether this assumption is necessary or it can be weakened.

2.3 Interpretation of Theorems 2.7 and 2.8

We discuss Theorems 2.7 and 2.8 in a few remarks below.

Remark 2.11 (Difference of upper and lower bounds)**.** Theorem 2.7 holds for any set $\mathcal{C}$ that satisfies polynomial complexity of order k. In contrast, the lower bound is obtained by taking the supremum of the minimax risk over all sets that satisfy polynomial complexity of order k. As Theorem 2.10 illustrates, due to the nature of the speckle noise, which is multiplied by the entries of $\boldsymbol{x}_o$, the estimation problem is easier for sparse vectors. Nevertheless, the supremum in the lower bound demonstrates that for certain sets $\mathcal{C}$ that satisfy polynomial complexity of order k, the upper bound established in Theorem 2.7 is in fact sharp.

Depending on the relative value of m, n, σ_z, the minimax risk can be obtained from one of the following formulas:

$$R_2(\mathcal{C}_k, m, n, \sigma_z) = \begin{cases} O_{x_{\max}, x_{\min}}\left(\frac{kn \log n}{m^2 L}\right), & \text{if } n \geq \max(m, \sigma_z^2); \\ O_{x_{\max}, x_{\min}}\left(\frac{k \log n}{nL}\right), & \text{if } m \geq \max(n, \sigma_z^2); \\ O_{x_{\max}, x_{\min}}\left(\frac{\sigma_z^4 k \log n}{m^2 L}\right), & \text{if } \sigma_z^2 \geq \max(m, n). \end{cases} \tag{2.8}$$

In what follows, we offer some intuition to explain these bounds.

Remark 2.12 (When does the multiplicative noise dominate the additive noise?)**.** Suppose that σ_z^2 is much smaller than $\max(m,n)$. In this case, the minimax risk is unaffected by the additive noise. To understand this phenomenon, suppose that we are working in the setting $m<n$. If we consider the model, $\boldsymbol{y}=AX_o\boldsymbol{w}+\boldsymbol{z}$, let $\boldsymbol{y}_i$ denote, the i^{th} element of $\boldsymbol{y}$, then we have $\boldsymbol{y}_i=\mathbf{a}_i^T X_o\boldsymbol{w}+\boldsymbol{z}_i$. In this measurement, the variance of $\text{var}(\mathbf{a}_i^T X_o\boldsymbol{w})$ is $\sum_{i=1}^n \boldsymbol{x}_i^2=\Theta(n)$, and $\text{var}(\boldsymbol{z}_i)=\sigma_z^2$. Hence, when $\sigma_z^2=o(n)$, one would expect the additive noise to be negligible. However, this heuristic does not explain why the additive noise does not matter when $m>n$ and $n\ll\sigma_z\ll m$. Again to provide some intuition on the impact of m, when $m>n$. Note that in this case, since A^TA is an invertible matrix, we can calculate: $\tilde{\boldsymbol{y}}=(A^TA)^{-1}\boldsymbol{y}=X\boldsymbol{w}+\tilde{\boldsymbol{z}}$, where $\tilde{\boldsymbol{z}}\sim N(0,\sigma_z^2(A^TA)^{-1})$. While the additive noise $\tilde{\boldsymbol{z}}$ is colored, and discussing signal-to-noise ratio on the individual elements does not necessarily provide an accurate information, note that in $\tilde{y}_i$ we have $\text{var}(x_iw_i)=\Theta(1)$, and we can prove that $\text{var}(\tilde{z}_i)=\Theta(\sigma_z^2/m)$ (See for instance definition 4.9 for the eigenvalues of A). Hence, in this case, again we can see that when $\sigma_z^2\ll m$, the additive noise becomes negligible.

Remark 2.13 (Comparison with *linear imaging systems*)**.** Again, consider the case $\sigma_z^2\le\max(m,n)$. In the classical regimes where the sparse linear regression problem is studied (e.g., Bickel et al. (2009)), namely $k\ll m\ll n$, the minimax risk of coherent imaging systems reduces to

$$\frac{kn\log n}{m^2L}.$$

If L is not too large, achieving a small risk requires $m\gg\sqrt{n}$. This contrasts sharply with imaging systems based on linear regression, where obtaining a small minimax risk typically requires only $m\gg k\log n$. This is consistent with the general belief in the coherent imaging community that recovering images from measurements in coherent imaging systems is much more challenging than in imaging systems based on linear models.

2.4 Fixed A model

In Section 2.2, we considered the setting in which the forward models A_is across looks are independent. However, as we discussed before, in some multilook systems the forward models do not change across looks and we have

$$A_1=A_2=\ldots=A_L.$$

The main question that we would like to address in this section is whether either of these two multilook systems have an advantage over each other. To respond to this question, we aim to study the minimax estimation rate under the setting $A_1=A_2=\ldots=A_L$ and compare it with the result of Section 2.2. We define the minimax risk for this setting similar to what we defined before for the vase of different forward models.

$$R_2^{\dagger}(\mathcal{C},m,n,\sigma_z)=\inf_{\widehat{\boldsymbol{x}}}\sup_{x_o\in\mathcal{C}}\mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_o\|_2^2}{n}\right]. \tag{2.9}$$

Theorem 2.14. Suppose that $A_1=A_2=\ldots=A_L=A\in\mathbb{R}^{m\times n}$ and that $A_{ij}\overset{i.i.d.}{\sim}N(0,1)$. Furthermore, assume that $\boldsymbol{x}_o\in\mathcal{C}_k\in\mathcal{F}_{a,b,k,n}$. Furthermore, assume $mL\le n^4k\log n$. Then,

$$R_2^{\dagger}(\mathcal{C}_k,m,n,\sigma_z)=O_{x_{\max},x_{\min}}\left(\min\left(\frac{\max(\sigma_z^4,m^2,n^2)k\log n}{m^2nL}+\frac{k\log m\log n}{m^2},1\right)\right). \tag{2.10}$$

In particular, if $\max(\sigma_z^4, m^2, n^2) \geq nL\log m$, then

$$R_2^\dagger(\mathcal{C}_k, m, n, \sigma_z) = O_{x_{\max}, x_{\min}}\left(\min\left(\frac{\max(\sigma_z^4, m^2, n^2)k\log n}{m^2 nL}, 1\right)\right). \tag{2.11}$$

Similar to Section 2.2, the above theorem obtains an upper bound for the minimax risk that holds for any $\mathcal{C} \subset [x_{\min}, x_{\max}]^n$ that satisfies polynomial complexity of order k. Before discussing the implications of the above result, let us discuss the sharpness of this upper bound.

Theorem 2.15. Suppose that the following holds: (i) $a \geq x_{\max} - x_{\min}, b \geq 1, \log m = \Theta(\log n), \log L = O(\log n)$, (ii) there exists $\varepsilon \in (0, 1/2)$ such that $k \leq n^{1-2\varepsilon}$, and (iii) $\max(\sigma_z^4, m^2, n^2)k\log n \leq m^2 n^{1-\varepsilon} L$. Then we have

$$\sup_{\mathcal{C} \in \mathcal{F}_{a,b,k,n}} R_2^\dagger(\mathcal{C}, m, n, \sigma_z) = \Omega_{\varepsilon, x_{\max}, x_{\min}}\left(\frac{\max(\sigma_z^4, m^2, n^2)k\log n}{m^2 nL}\right). \tag{2.12}$$

Since the assumptions in the above two theorems are similar to the ones in Theorems 2.7 and 2.8 we will not discuss the assumptions again. However, there is one condition that we did not discuss before. This condition appears in the second part of Theorem 2.14 and indicates that if $\max(\sigma_z^4, m^2, n^2) \geq nL\log m$, then

$$R_2^\dagger(\mathcal{C}, m, n, \sigma_z) = O_{x_{\max}, x_{\min}}\left(\min\left(\frac{\max(\sigma_z^4, m^2, n^2)k\log n}{m^2 nL}, 1\right)\right). \tag{2.13}$$

Note that the upper bound in Definition 2.14 matches the lower bound in Definition 2.15. Hence, this leads to two questions:

1. How strong is the assumption $\max(\sigma_z^4, m^2, n^2) \geq nL\log m$?
2. Is there any intuition why $\frac{k\log m\log n}{m^2}$ appears in Definition 2.14 while it does not appear in Theorem 2.13?

In response to the first question above, let us assume that $\max(\sigma_z^4, m^2, n^2) = n^2$. Then, the condition $\max(\sigma_z^4, m^2, n^2) \geq nL\log m$ simplifies to $n \geq L\log m$. In practice, $n \gg \log m$ and L is often a number between 2 to 100, and the condition holds. A similar argument shows that even when $m > n$, the condition is often satisfied.

Regarding the second question raised above, the necessity of $\frac{k\log m\log n}{m^2}$ is still unclear. However, some intuitive arguments shed some light on the difference between the fixed-A and varying-A cases. Suppose that the additive noise in (2.2) is zero. In the fixed design setting, it is straightforward to show that the statistics $\frac{1}{L}\sum_{i=1}^{L} \boldsymbol{y}_l \boldsymbol{y}_l^T$ is a sufficient statistics for X_o. if we fix m, n and let $L \to \infty$, the sufficient statistics converges to $AX_o^2A^T$ in probability. In other words, the sufficient statistics converges to a linear transformation of X_o^2. However, note that recovering the exact X_o^2 from $AX_o^2A^T$ is not possible. In fact, in the most optimistic setting $AX_0^2A^T$ offers $m(m+1)/2$ linearly independent observations of X_o^2. Hence, if $m(m+1)/2 < k$, we cannot recover the exact X_o^2 from $AX_o^2A^T$. Note that when $k > m^2$, the term $\frac{k\log m\log n}{m^2}$ is quite large. Hence, this term is consistent with the intuition that when $k \geq m^2$ the error has to be large.

2.5 Fixed forward model or varying forward model?

By comparing the results of Sections 2.2 and 2.4, we observe that, in terms of minimax rates, there is no significant difference between fixed and varying forward models with respect to estimation accuracy. The upper bound that we have derived for the the fixed-A model, has an extra $\frac{k \log m \log n}{m^2}$. However, this extra term is quite small for most practical settings and does not seem to be important.

3 Related works

In this paper, we make the first attempt to establish the rate of the minimax risk for coherent imaging systems.

There is a substantial body of research on the theoretical characterization of the minimax risk for other imaging systems, such as MRI and CT, which are modeled by linear regression problems Tsybakov (1986); Korostelev and Tsybakov (1993); Donoho and Johnstone (1994b); Bickel et al. (2009); Raskutti et al. (2011); Candès and Su (2015); Su et al. (2017); Weng et al. (2016); Donoho et al. (2009); Metzler et al. (2016); Guo et al. (2024); Ghosh et al. (2025), as well as crystallography and astrophotography, which are modeled by the phase retrieval problem Chen et al. (2019); Chen and Candès (2017); Candes et al. (2015); Zhang et al. (2017, 2016); Cai et al. (2016); Hand et al. (2018); Zhang et al. (2017); Ma et al. (2019); Bakhshizadeh et al. (2020). However, due to the presence of multiplicative noise, our proof strategies and resulting characterizations differ significantly. For example, as discussed in Sections 2.1 and 2.2, the sparsity assumption that is central to much of the theoretical work on these other imaging systems is not particularly useful for analyzing coherent imaging systems. Consequently, we were required to consider a broader class of signals, i.e., those that have polynomial complexity of order k. Moreover, because of the fundamental differences in the underlying mathematical models, both our proof strategies and the analytical tools we employ are considerably different from those commonly used in the literature on sparse linear regression and sparse phase retrieval. For instance, the standard strategy in sparse linear regression is to assume some condition on matrix A (called the forward operator of the imaging system) such as restricted isometry property Candes and Tao (2005), compatibility condition van de Geer and Bühlmann (2009), restricted eigenvalue (RE) condition Bickel et al. (2009), and strong restricted eigenvalue (SRE) condition Bellec et al. (2018), and later confirm them on a given random matrix ensemble. As is clear, since we do not have the assumption of sparsity and the speckle noises are multiplied by the elements of vectors such conditions are not useful in our proofs.

The theoretical properties of coherent imaging systems have only recently been explored in a few papers by subsets of the authors and their collaborators Zhou et al. (2022); Chen et al. (2024, 2025); Malekian et al. (2025). We discuss the contributions of these papers, and compare our contributions with what are offered in those papers:

1. Speckle noise in nonparametric settings: In Malekian et al. (2025) the authors study the speckle noise under the nonparametric settings. More specifically, they consider the model:

$$y_i = f(x_i)\xi_i + \tau_i, \quad i = 1, 2, \ldots, n. \tag{3.1}$$

where ξ_i's are i.i.d. $\mathcal{N}(0,1)$ and τ_i's are i.i.d. $\mathcal{N}(0,\sigma_\tau^2)$ random variables, $x_i = i/n, i = 1, 2, ..., n$ are fixed design points, and unknown f is a smooth function assumed to be in a

Holder class $\mathscr{S}$. Then, the authors characterized the minimax risk:

$$R_2(\mathscr{S}, \sigma_\tau) := \inf_{\hat{f}} \sup_{f \in \mathscr{S}} \mathbb{E}_f \|f - \hat{f}\|_2^2, \tag{3.2}$$

where $\mathscr{S}$ denotes a Holder class of function. Note that this problem reduces to the standard problem of nonparametric regression when ξ_i is equal to 1, on which a large body of work exists in the literature. Tsybakov (1986); Korostelev and Tsybakov (1993); Arias-Castro et al. (2012); Maleki et al. (2012, 2013); Kerkyacharian and Picard (1992); Donoho and Johnstone (1998b); Donoho (1999); DeVore et al. (2025).

Compared to our paper, we should emphasize that Malekian et al. (2025) has assumed that the forward operator A is given by I. As expected, many complications in our derivations arise because of existence the forward operator in our model. Hence, we need completely different techniques (and different algorithms for obtaining upper bounds) from the ones presented in Malekian et al. (2025).

2. Fixed forward models: The authors of Zhou et al. (2022); Chen et al. (2024, 2025) have studied a problem similar to the one presented in Section 2.4. However, there are several major differences between their work and ours.

   (a) None of these three papers establish lower bounds for the minimax risk. As will be clarified later, one of the main technical contributions of this paper has been to develop lower bounds for different regimes. For example, $m < n$ versus $m > n$, require distinct lower-bounding techniques. Studying exactly sparse signals again requires new techniques. In addition, for studying the lower bound in the singular case $m > n$ (i.e., the number of sensors m is larger than the dimension n of the signal $\boldsymbol{x}_o$, which forces the data $\boldsymbol{y}_l \in \mathbb{R}^m$ to lie in a significantly higher dimensional space) we introduce novel ideas involving the theory of Rao-Blackwell theorem and sufficient statistics, to achieve matching lower bounds.

   (b) In their models, Zhou et al. (2022); Chen et al. (2024, 2025) did not account for varying measurement scenarios. They also studied the ideal setting where the additive noise was set to zero, and imposed the assumption $m < n$; in fact, Zhou et al. (2022) further required $m = \Theta(n)$. As our proofs will demonstrate, relaxing each of these assumptions and deriving sharp bounds necessitate new technical contributions. For example, our proof strategy for the case $m > n$ is fundamentally different from that for $m < n$.

Prior to this work, coarser high probability upper bounds for $\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2}{n}$ were obtained in Zhou et al. (2022) for the single-look speckle noise model where the signal class comes from a structured compression codebook, and Chen et al. (2024, 2025) for multi-look unvarying measurement speckle noise model where the signals are considered as images of a (bi-)Lipchitz function. In addition, the main theorems in these works assume the undersample regime $m \ll n$ and additive noise $\sigma_z = 0$. This paper overcomes all these limitations.

Another classical problem of study is nonparametric function recovery. Consider the regression model

$$y_i = f(x_i) + \tau_i, \quad i = 1, 2, \ldots, n, \tag{3.3}$$

where f is the unknown function from a non-parametrized functional space $\mathscr{S}$ and τ_i's are random noises. x_i can be either fixed or random design points. One can study the minimax risk, for example,

$$R_2(\mathscr{S}, \sigma_\tau) := \inf_{\hat{f}} \sup_{f \in \Theta} \mathbb{E}_f \|f - \hat{f}\|_2^2 \tag{3.4}$$

The classical subjects of study for $\mathscr{S}$ include Hölder classes Tsybakov (1986); Korostelev and Tsybakov (1993); Arias-Castro et al. (2012); Maleki et al. (2012, 2013), Soblev classes Nemirovskii (1985); Nemirovskii A.S. and Tsybakov (1985), and Besov classes Kerkyacharian and Picard (1992); Donoho and Johnstone (1998b); Donoho (1999); DeVore et al. (2025). Recently, Malekian et al. (2025) studied the minimax risk under the speckle noise model

$$y_i = f(x_i)\xi_i + \tau_i, \quad i = 1, 2, \ldots, n. \tag{3.5}$$

where ξ_i's are i.i.d. $\mathcal{N}(0, 1)$ and τ_i's are i.i.d. $\mathcal{N}(0, \sigma_n^2)$ random variables, $x_i = i/n, i = 1, 2, ..., n$ are fixed design points, and $\mathscr{S}$ is the space of functions with uniform upper and lower bounds in a Hölder class.

Moreover, we treat both undersample ($m \leq n$) and oversample ($m \geq n$) regimes and demonstrate how m and n determine the thresholds with respect to which the noise level σ_z from $\boldsymbol{z}_1, ..., \boldsymbol{z}_L$ can affect the minimax rates. This provides a complete picture to the minimax error estimation of this problem.

4 Preliminaries

In this section we summarize technical results used in this paper, see Appendix C for proofs.

4.1 Results regarding the minimax risk

In the proofs of some our main results we will need some of the basic monotonicity properties of the minimax risk. While such results are intuitive and well-known on simpler problems such as in the estimation of the mean of a Gaussian random vector Donoho and Johnstone (1994a), for completeness, we prove them for the estimation problems we discuss in this paper.

Our first lemma suggests that increasing the number of observations m makes the statistical problem only easier.

Lemma 4.1. $R_2(\mathcal{C}, m, n, \sigma_z)$ as defined in (2.3) and $R_2^\dagger(\mathcal{C}, m, n, \sigma_z)$ as defined in (2.9) are non-increasing in m.

Our next lemma confirms that increasing the variance of the additive noise only makes the estimation problem harder.

Lemma 4.2. $R_2(\mathcal{C}, m, n, \sigma_z)$ as defined in (2.3) and $R_2^\dagger(\mathcal{C}, m, n, \sigma_z)$ as defined in (2.9) are non-decreasing in σ_z.

We use the following version of Fano's method to obtain the lower bounds for the minimax risk:

Lemma 4.3 (Generalized Fano method, Lemma 3, Yu (1997))**.** Let $\mathcal{P}$ be a space of probability measures such that for each $\mathbb{P} \in \mathcal{P}$, there is an associated parameter $\theta(\mathbb{P})$ of interest. Let d be a pseudo-metric on the space $\theta(\mathcal{P})$. Suppose there exists an integer $r \geq 2$ and parameters α_r and β_r satisfying

1. $\{\theta(\mathbb{P}_1), ..., \theta(\mathbb{P}_r)\}$ is an α_r-separated subset in $(\theta(\mathcal{P}), d)$, namely for all $1 \leq i \neq j \leq r$,

$$d(\theta(\mathbb{P}_i), \theta(\mathbb{P}_j)) \geq \alpha_r.$$

2. For all $1 \leq i \neq j \leq r$, we have the upper bound for Kullback–Leibler divergence

$$\mathrm{KL}(\mathbb{P}_i \parallel \mathbb{P}_j) := \int \log(\mathbb{P}_i / \mathbb{P}_j) d\mathbb{P}_i \leq \beta_r$$

Then for any $\widehat{\theta} \in \theta(\mathcal{P})$, we have the lower bound estimate

$$\max_{1 \leq j \leq r} \mathbb{E}_{\mathbb{P}_j} \left[d\left(\widehat{\theta}, \theta(\mathbb{P}_j)\right) \right] \geq \frac{\alpha_r}{2} \left(1 - \frac{\beta_r + \log 2}{\log r}\right).$$

As is clear from the above theorem in order to use Fano's inequality, we have to find an upper bound for the KL divergence of two distributions. One of the results that will be used in our paper is the following well-knonw result on the KL divergence of two Gaussian distributions:

Proposition 4.4. (Duchi, 2007) If $\mathbb{Q}_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j), j = 1, 2$, are two d-dimensional multivariate normal distributions, then

$$\mathrm{KL}(\mathbb{Q}_1 \parallel \mathbb{Q}_2) = \frac{1}{2} \left[\log \frac{\det \boldsymbol{\Lambda}_2}{\det \boldsymbol{\Lambda}_1} - d + \mathrm{Tr}\left(\boldsymbol{\Lambda}_2^{-1} \boldsymbol{\Lambda}_1\right) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Lambda}_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \right]. \tag{4.1}$$

4.2 Results on covering numbers

Definition 4.5. A *α-separating* subset of S is a finite or countable collection $\{\boldsymbol{x}_i\}$ of points of X satisfying $\mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq \alpha$ for any $i \neq j$. We call the largest possible cardinality among all α-separated subsets of S the *α-packing number* of S, denoted $P_\alpha(S)$. In other words,

$$P_\delta(S) := \sup\{n : \text{there exists a } \alpha\text{-separated subset of } S \text{ of cardinality } n\}. \tag{4.2}$$

Proposition 4.6 ((Vershynin, 2018, Proposition 4.2.12))**.** For any Euclidean ball $B_R \subset \mathbb{R}^n$ of radius $R > 0$ (in ℓ_2 norm), we have an estimate for its δ-covering number as follows:

$$\left(\frac{R}{\delta}\right)^n \leq N_\delta(B_R) \leq \left(\frac{2R}{\delta} + 1\right)^n. \tag{4.3}$$

4.3 Concentration and decoupling results

We first start with a decoupling result that will play critical role in our paper:

Lemma 4.7. (De la Pena and Giné, 2012, Theorem 3.4.1.) Let $X_1, X_2, \ldots, X_n$ denote random variables with values in measurable space $(S, \mathcal{S})$. Let $(\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n)$ denote an independent copy of $X_1, X_2, \ldots, X_n$. For $i \neq j$ let $h_{i,j} : S^2 \to \mathbb{R}$. Then, there exists a constant C such that for every $t > 0$ we have

$$\mathbb{P}\left(\Big|\sum_{i \neq j} h_{i,j}(X_i, X_j)\Big| > t\right) \leq C\mathbb{P}\left(C\Big|\sum_{i \neq j} h_{i,j}(X_i, \widetilde{X}_j)\Big| > t\right).$$

One of the concentration results that will be used extensively in our paper is the concentration of quadratic functions, a.k.a. Hanson-Wright inequality.

Lemma 4.8 (Hanson-Wright inequality, Hanson and Wright (1971))**.** Let $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^\top$ be a random vector with independent components with $\mathbb{E}[\xi_i] = 0$ and $\|\xi_i\|_{\text{subgau}} \leq K$. Let A be an $n \times n$ matrix. Then, for $t > 0$,

$$\mathbb{P}\left(\left[\left|\boldsymbol{\xi}^\top A\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^\top A\boldsymbol{\xi}]\right| > t\right]\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{K^4\|A\|_{\text{HS}}^2}, \frac{t}{K^2\|A\|_2}\right)\right), \tag{4.4}$$

where c is a constant, and $\|\xi_i\|_{\text{subgau}} = \inf\{t > 0 : \mathbb{E}(\exp(\xi_i^2/t^2)) \leq 2\}$.

We will use the following classical results on random matrices throughout our proofs:

Lemma 4.9. (Rudelson and Vershynin, 2010, Theorem 2.6) and (Davidson and Szarek, 2001, Theorem II.13) Let A be an $m \times n$ random matrix with elements drawn i.i.d. from $N(0, 1)$. Then,,

$$\mathbb{P}(\sigma_{\max}(A) \leq \sqrt{n} + \sqrt{m} + t) \geq 1 - 2\mathrm{e}^{-\frac{t^2}{2}}, \quad t > 0.$$

Moreover, if $m < n$, then for any $t > 0$,

$$\mathbb{P}(\sqrt{n} - \sqrt{m} - t \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq \sqrt{n} + \sqrt{m} + t) \geq 1 - 2\mathrm{e}^{-\frac{t^2}{2}}.$$

Throughout the proof for the varying forward operators, we will use this result in the following way. For i.i.d. $m \times n$, $N(0, 1)$ Gaussian matrices $A_1, ..., A_L$, we define the event

$$\mathcal{E}_{\text{maxsing}}(t, L) := \bigcap_{l=1}^{L} \left\{\sigma_{\max}(A_l) \leq \sqrt{n} + \sqrt{m} + t\right\}. \tag{4.5}$$

With a slight overloading of our notation we also define:

$$\mathcal{E}_{\text{maxsing}} := \mathcal{E}_{\text{maxsing}}\left(\frac{\sqrt{n} + \sqrt{m}}{2}, L\right). \tag{4.6}$$

If $n \geq 4m$, we define the event

$$\mathcal{E}_{\text{sing}}(t, L) = \bigcap_{l=1}^{L} \left\{\sqrt{n} - \sqrt{m} - t \leq \sigma_{\min}(A_l) \leq \sigma_{\max}(A_l) \leq \sqrt{n} + \sqrt{m} + t\right\}, \tag{4.7}$$

and again define a slightly overloaded notation:

$$\mathcal{E}_{\text{sing}} := \mathcal{E}_{\text{sing}}\left(\frac{\sqrt{n} - \sqrt{m}}{2}, L\right), \quad \mathbb{P}[\mathcal{E}_{\text{sing}}] \geq 1 - 2L\exp\left(-\frac{(\sqrt{n} - \sqrt{m})^2}{8}\right), \tag{4.8}$$

where the last inequality followed from Lemma 4.9. Similarly, for $m \geq 4n$, we define the event

$$\mathcal{E}'_{\text{sing}}(t, L) = \bigcap_{l=1}^{L} \left\{\sqrt{m} - \sqrt{n} - t \leq \sigma_{\min}(A_l^\top) \leq \sigma_{\max}(A_l^\top) \leq \sqrt{m} + \sqrt{n} + t\right\}. \tag{4.9}$$

For the special case when $t = \frac{\sqrt{m}-\sqrt{n}}{2}$, we again use Lemma 4.9 to get

$$\mathcal{E}'_{\text{sing}} := \mathcal{E}'_{\text{sing}}\left(\frac{\sqrt{m}-\sqrt{n}}{2}, L\right), \quad \mathbb{P}[\mathcal{E}'_{\text{sing}}] \geq 1 - 2L\exp\left(-\frac{(\sqrt{n}-\sqrt{m})^2}{8}\right). \tag{4.10}$$

The following lemma, proved in Subsection C.3, is a generalization and more accurate version of Lemma 4 and 5 from Zhou et al. (2024):

Lemma 4.10. Let $\{A_1\}_{l=1}^L \in \mathbb{R}^{m\times n}$ be Gaussian matrices. For any fixed $\boldsymbol{d} \in \mathbb{R}^n$, define $D = \operatorname{diag}(\boldsymbol{d})$. Define the event $\widetilde{\mathcal{E}}_{\text{maxsing}} := \bigcap_{l=1}^L \bigcap_{i=1}^m \left\{\sigma_{\max}(\tilde{A}_{l,\setminus i}) \leq \frac{3}{2}(\sqrt{m}+\sqrt{n})\right\}$ where $\tilde{A}_{l,\setminus i}$ is an independent copy of A_l with i-th row removed. Then $\mathbb{P}(\widetilde{\mathcal{E}}_{\text{maxsing}}) \geq 1 - 2mL\exp(-cn)$ and we have

1. The upper tail probability

$$\begin{aligned}
&\mathbb{P}\left(\left[\sum_{l=1}^L \|A_l D A_l^\top\|_{\text{HS}}^2 > Lm\left(\operatorname{Tr}(D)+t_1\right)^2 + Lm(m-1)\|\boldsymbol{d}\|_2^2 + t_2\right] \cap \widetilde{\mathcal{E}}_{\text{maxsing}}\right) \\
&\leq 2mL\exp\left(-c\min\left(\frac{t_1^2}{K^4\|\boldsymbol{d}\|_2^2}, \frac{t_1}{K^2\|\boldsymbol{d}\|_\infty}\right)\right) + 2m\exp\left(-c\min\left(\frac{t_2^2}{K^4Lm^3\|\boldsymbol{d}^2\|_2^2}, \frac{t_2}{K^2m\|\boldsymbol{d}^2\|_\infty}\right)\right) \\
&+ 2C\exp\left(-c\min\left(\frac{4t_2^2}{81C^2K^4\|\boldsymbol{d}^2\|_2^2 mL(\sqrt{n}+\sqrt{m})^4}, \frac{2t_2}{9CK^2\|\boldsymbol{d}\|_\infty^2(\sqrt{n}+\sqrt{m})^2}\right)\right).
\end{aligned} \tag{4.11}$$

2. The lower tail probability

$$\begin{aligned}
&\mathbb{P}\left(\left[\sum_{l=1}^L \|A_l D A_l^\top\|_{\text{HS}}^2 < Lm(m-1)\|\boldsymbol{d}\|_2^2 - t\right] \cap \widetilde{\mathcal{E}}_{\text{maxsing}}\right) \\
&\leq 2C\exp\left(-c\min\left(\frac{4t^2}{81C^2K^4\|\boldsymbol{d}^2\|_2^2 mL(\sqrt{n}+\sqrt{m})^4}, \frac{2t}{9CK^2\|\boldsymbol{d}\|_\infty^2(\sqrt{n}+\sqrt{m})^2}\right)\right) \\
&+ 2m\exp\left(-c\min\left(\frac{t^2}{K^4Lm^3\|\boldsymbol{d}^2\|_2^2}, \frac{t}{K^2m\|\boldsymbol{d}^2\|_\infty}\right)\right).
\end{aligned} \tag{4.12}$$

where C and c are absolute constants. Here $1 \leq K \leq 2$ denotes the subgaussian norm of a standard Gaussian random variable.

4.4 Linear algebraic results

The following simple linear algebraic result will help us in bounding the differences between the inverse of two matrices.

Lemma 4.11. (Chen et al., 2024, Lemma 6.1) Let $B, C \in \mathbb{R}^{n\times n}$ be symmetric, invertible matrices. Then $\left\|B^{-1}-C^{-1}\right\|_2 \le \sigma_{\max}\left(B^{-1}-C^{-1}\right) \le \frac{\sigma_{\max}(B-C)}{\sigma_{\min}(B)\sigma_{\min}(C)}$.

Lemma 4.12. Let A denote an arbitrary matrix and D be a diagonal matrix. Then we have,

$$\|AD\|_{\mathrm{HS}} \le \sigma_{\max}(A)\|D\|_{\mathrm{HS}}.$$

Lemma 4.13 (Von Neumann's trace inequality)**.** (Horn and Johnson, 2012, Theorem 7.4.1.1)
Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ denote two matrices with singular values σ_i^A and σ_i^B. Then, we have

$$\mathrm{Tr}(AB) \le \sum \sigma_i^A \sigma_i^B.$$

The following theorem is a generalization of (Zhou et al., 2022, Lemma 5).

Lemma 4.14. Denote $\widetilde{\Sigma} = (\sigma_z^2 I_m + A\widetilde{X}^2A^\top)^{-1}, \Sigma = (\sigma_z^2 I_m + AX^2A^\top)^{-1}$. Then,

$$\begin{aligned}\left\|A(\widetilde{X}^2 - X^2)A^\top\right\|_{\mathrm{HS}}^2 \frac{\left(\sigma_z^2 + x_{\min}^2\lambda_{\min}(AA^\top)\right)^2}{\left(\sigma_z^2 + x_{\max}^2\lambda_{\max}(AA^\top)\right)^4} &\le \mathrm{Tr}\left[(\Sigma^{-1}(\widetilde{\Sigma}-\Sigma)\Sigma^{-1}(\widetilde{\Sigma}-\Sigma)\right] \\ &\le \frac{\left(\sigma_z^2 + x_{\max}^2\lambda_{\max}(AA^\top)\right)^2}{\left(\sigma_z^2 + x_{\min}^2\lambda_{\min}(AA^\top)\right)^4}\left\|A(\widetilde{X}^2 - X^2)A^\top\right\|_{\mathrm{HS}}^2.\end{aligned}$$

5 Proof of the lower bound in the case $m < \frac{n}{4}$

5.1 Outline of the proof strategy

As increasing a and b only makes $\mathcal{F}_{a,b,k,n}$ larger, it suffices to prove Definition 2.8 for $a = x_{\max} - x_{\min}$ and $b = 1$. We will apply Definition 4.3 with the following definitions to obtain the minimax lower bound in Definition 2.7. Given $\boldsymbol{x} \in \mathbb{R}^n$, let $\mathbb{P}_{\boldsymbol{x}}$ denote the probability distribution of data $\vec{\boldsymbol{y}} = [\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_L^\top]^\top$ generated according to the model (2.2) with $X = \mathrm{diag}(\boldsymbol{x})$. More specifically, we choose:

$$\begin{gathered}\mathbb{P}_{\boldsymbol{x}} \sim \otimes_{l=1}^L N(\mathbf{0}, \Sigma_l^{-1}(\boldsymbol{x})) = N(\mathbf{0}, \boldsymbol{\Sigma}^{-1}(\boldsymbol{x})), \\ \boldsymbol{\Sigma}(\boldsymbol{x}) := \mathrm{diag}\left(\Sigma_1(\boldsymbol{x}), ..., \Sigma_L(\boldsymbol{x})\right), \quad \Sigma_l = \Sigma_l(\boldsymbol{x}) := (\sigma_z^2 I_m + A_l X^2 A_l^\top)^{-1}.\end{gathered} \tag{5.1}$$

Then, the parameter $\theta(\mathbb{P}(\boldsymbol{x}))$ corresponding to the distribution $\mathbb{P}(\boldsymbol{x})$ is chosen as

$$\theta(\mathbb{P}(\boldsymbol{x})) = \boldsymbol{x}, \quad d(\theta(\mathbb{P}_{\boldsymbol{x}}), \theta(\mathbb{P}_{\boldsymbol{x}'})) = d(\boldsymbol{x}, \boldsymbol{x}') := \|\boldsymbol{x} - \boldsymbol{x}'\|_2, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n, \tag{5.2}$$

and we select a subset $\mathcal{C}$ of the parameters that is large enough to provide us with the desired complexity. In order to apply Lemma 4.3 we identify a discretization $\mathcal{S}_{\mathrm{sep}} \subset \mathcal{C}, |\mathcal{S}_{\mathrm{sep}}| = r$ satisfying:

(P1) For any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{S}_{\mathrm{sep}}$ with $\boldsymbol{x}_i \neq \boldsymbol{x}_j$, the corresponding parameters $\theta(\mathbb{P}_{\boldsymbol{x}_i}), \theta(\mathbb{P}_{\boldsymbol{x}_j})$ (which are identically $\boldsymbol{x}_i, \boldsymbol{x}_j$ according to our definition) are well separated. In particular, for an α_r to be chosen appropriately, we will establish that

$$d(\theta(\mathbb{P}_{\boldsymbol{x}_i}), \theta(\mathbb{P}_{\boldsymbol{x}_j})) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \ge \alpha_r, \quad \boldsymbol{x}_i \neq \boldsymbol{x}_j \in \mathcal{S}_{\mathrm{sep}}.$$

(P2) For all $1 \le i \ne j \le r$, for a $\beta_r > 0$ to be chosen later, the distributions $\mathbb{P}_{\boldsymbol{x}_i}, \mathbb{P}_{\boldsymbol{x}_j}$ are difficult to distinguish in the Kullback-Leibler divergence, at the level β_r

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}_i} \parallel \mathbb{P}_{\boldsymbol{x}_j}) \le \beta_r, \quad \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{S}_{\text{sep}}. \tag{5.3}$$

Then, Lemma 4.3 directly implies that there exists a constant $C > 0$ for which

$$\inf_{\widehat{\boldsymbol{x}}} \sup_{1 \le j \le r} \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_j\|^2}{n}\right] \ge \frac{C\alpha_r^2}{n}\left(1 - \frac{\beta_r + \log 2}{\log r}\right)^2. \tag{5.4}$$

In particular, for some small $c > 0$, we will end up making the following choice for r, α_r, β_r

$$\log r = \Theta(k \log n), \quad \alpha_r = \Theta_{\varepsilon, x_{\max}, x_{\min}}\left(\frac{\max(\sigma_z^4, n^2) k \log n}{m^2 L}\right), \quad \beta_r = c \log r.$$

As a consequence, by Definition 4.3 the following lower bounds holds for any estimator $\widehat{\boldsymbol{x}}$

$$\max_{1 \le i \le r} \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_i\|^2}{n}\right] \ge \frac{\alpha_r^2}{4n}\left(1 - \frac{\beta_r + \log 2}{\log r}\right)^2 = \Theta_{\varepsilon, x_{\max}, x_{\min}}\left(\frac{\max(\sigma_z^4, n^2) k \log n}{m^2 n L}\right).$$

Our proof strategy in the following sections will describe a construction which will dictate, with a high probability, the above choices for r, α_r, β_r. All the technical proofs are provided in Appendix D.

5.2 Construction of the signal class

We will first construct the signal class $\mathcal{C}$. Fix $0 < x_{\min} < x_{\max}$. For any array of nonnegative integers $0 = a_0 < a_1 < \cdots < a_k = n$, let $\{(a_{l-1}, a_l]_{\mathbb{Z}} : l \in [k]\}$ be an ordered k-partition of $[n]$.

$$[n] = \cup_{l=1}^k (a_{l-1}, a_l]_{\mathbb{Z}}. \tag{5.5}$$

Let $\mathcal{F}(a_0, ..., a_k)$ denote the set of functions from $[n]$ to $[x_{\min}, x_{\max}]$ that are constant on each integer interval $(a_{l-1}, a_l]_{\mathbb{Z}}, l = 1, 2, ..., l$. Define

$$\mathcal{F}_k := \bigcup_{0 = a_0 < a_1 < \cdots < a_k = n} \mathcal{F}(a_0, ..., a_k), \;\; \mathcal{X}_k := \{(f(1), ..., f(n)) \in \mathbb{R}^n : f \in \mathcal{F}_k\} \subset [x_{\min}, x_{\max}]^n. \tag{5.6}$$

$\mathcal{X}_k$ satisfies the polynomial complexity of order k defined in definition 2.1. To see this, first note that $\mathcal{X}_k$ can be written as a finite union:

$$\mathcal{X}_k = \bigcup_{0 = a_0 < a_1 < \cdots < a_k = n} \{(f(1), ..., f(n)) \in \mathbb{R}^n : f \in \mathcal{F}(a_0, ..., a_k)\}. \tag{5.7}$$

We have

$$N_\varepsilon(\mathcal{X}_k) \le \sum_{0 = a_0 < a_1 < \cdots < a_k = n} \left(\frac{x_{\max} - x_{\min}}{\varepsilon}\right)^k = \binom{n}{k}\left(\frac{x_{\max} - x_{\min}}{\varepsilon}\right)^k \tag{5.8}$$

$$\overset{(a)}{\le} \left(\frac{n(x_{\max} - x_{\min})}{\varepsilon}\right)^k \tag{5.9}$$

where for (a) we used $\binom{n}{k} \leq n^k$. Hence $\mathcal{X}_k$ satisfies (2.1).

Intuitively, the class of signals we have considered are 'piecewise constant'. This is a natural and popular choice for class of images in image processing Rudin et al. (1992); Jalali and Maleki (2016); Donoho (1999). Definition 2.5 shows that this set satisfies the polynomial complexity of order k.

We now pick *signal class* $\mathcal{C}$ as any subset $[x_{\min}, x_{\max}]^n$ which is a superset of $\mathcal{X}_k$ and satisfies the polynomial complexity of order k. Hence

$$\mathcal{X}_k \subset \mathcal{C} \subset [x_{\min}, x_{\max}]^n.$$

Given this choice of $\mathcal{C}$ we now would like to show that

$$R_2(\mathcal{C}, m, n, \sigma_z) = \Omega_{\varepsilon, x_{\max}, x_{\min}} \left(\frac{\max(\sigma_z^4, m^2, n^2) k \log n}{m^2 n L} \right).$$

5.3 Discretization of the signal class to apply Fano's Lemma

To construct an α_r-separated subset $\mathcal{S}_{\text{sep}}$, we first define $\mathcal{X}^{\text{finite}} \subset \mathcal{C}$ as follows. For $\varepsilon \in (0,1)$, denote

$$N_{\text{div}} := kn^{\varepsilon}. \tag{5.10}$$

For simplicity, assume that both N_{div} and n/N_{div} are integers. We partition $[n]$ into N_{div} pieces as

$$[n] = \bigcup_{l=1}^{N_{\text{div}}} \left(\frac{(l-1)n}{N_{\text{div}}}, \frac{ln}{N_{\text{div}}} \right]_{\mathbb{Z}}$$

Note that each $j \in [n]$ (corresponding to subscripts of the coordinates of $\boldsymbol{x}$) will fall into one of the intervals $\left(\frac{(l_j-1)n}{N_{\text{div}}}, \frac{l_j n}{N_{\text{div}}} \right]_{\mathbb{Z}}$. Each integer interval contains $\Theta(n/N_{\text{div}})$ indices of $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

Fix $0 < \delta_r < \frac{x_{\max} - x_{\min}}{2}$ to be determined later and define $\bar{x} = \frac{x_{\min} + x_{\max}}{2}$. Let $\mathcal{B}_{N_{\text{div}}}$ denote the collection of all functions from $[n]$ to $\{\bar{x}, \bar{x} + \delta_r\}$, that are piecewise constant on each $\left(\frac{(l-1)n}{N_{\text{div}}}, \frac{ln}{N_{\text{div}}} \right]_{\mathbb{Z}}, l = 1, 2, ..., N_{\text{div}}$. We define

$$\mathcal{B}^{\text{finite}} := \left\{ (f(1), ..., f(n)) : f \in \mathcal{B}_{N_{\text{div}}} \right\} \cap \mathcal{C} \subset \{\bar{x}, \bar{x} + \delta_r\}^n. \tag{5.11}$$

We construct $\mathcal{X}^{\text{finite}} \subset \mathcal{B}^{\text{finite}}$ as follows: among the N_{div} intervals, select $k/2$ of them; for the entries of the vector corresponding to these intervals, assign the value $\bar{x} + \delta_r$, and set all remaining entries to $\bar{x}$. It is straightforward to see that

$$|\mathcal{X}^{\text{finite}}| = \binom{N_{\text{div}}}{k/2}. \tag{5.12}$$

Now we construct a subset $\mathcal{S}_{\text{sep}}$ of $\mathcal{X}^{\text{finite}}$ that satisfies α_r-separation condition (P1) as follows.

- Set $k' := k/4$ and let $\mathcal{S}_{\text{sep}}$ denote the set of all vectors in $\mathcal{X}^{\text{finite}}$ with the following property: If $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are in $\mathcal{S}_{\text{sep}}$ then, their q-th components satisfy $x_{i,q} \neq x_{j,q}$ for all q in at least k'-many different intervals of the form $\left(\frac{(l-1)n}{N_{\text{div}}}, \frac{ln}{N_{\text{div}}} \right]_{\mathbb{Z}}$.

The set $\mathcal{S}_{\rm sep}$ is the set of hypothesis that we use in Fano's Theorem. The following guarantees hold.

Cardinality of $\mathcal{S}_{\rm sep}$: The following lemma obtains an upper and a lower bound for $|\mathcal{S}_{\rm sep}|$.

Lemma 5.1. Let $r := |\mathcal{S}_{\rm sep}|$ denote the cardinality of the set $\mathcal{S}_{\rm sep}$. Then

$$\left(\frac{cN_{\rm div}}{k}\right)^{c'k} \leq r \leq \left(\frac{Cn}{k}\right)^{C'k}$$

for some absolute constants $c, C, c', C' > 0$. Consequently, $\log r = \Theta(k \log n)$.

Minimum separation of elements in $\mathcal{S}_{\rm sep}$: As the number of integer points in each interval $\left(\frac{(l-1)n}{N_{\rm div}}, \frac{ln}{N_{\rm div}}\right]_{\mathbb{Z}}$ is bounded below by $\frac{n}{N_{\rm div}} - 2$, we have $\min_{\substack{\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{S} \\ \boldsymbol{x}_i\neq\boldsymbol{x}_j}} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 \geq k' \cdot \left(\frac{n}{N_{\rm div}} - 2\right) \cdot \delta_r^2$. In view of the above, for a small constant $c > 0$, we choose α_r as

$$\alpha_r^2 = \frac{ckn\delta_r^2}{N_{\rm div}}. \tag{5.13}$$

Uniform signal strengths for elements in $\mathcal{S}_{\rm sep}$: Consider $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{S}_{\rm sep}$. Suppose that X_i, X_j denote diagonal square matrices $\mathrm{diag}(\boldsymbol{x}_i), \mathrm{diag}(\boldsymbol{x}_j)$ respectively. Since $\frac{n}{N_{\rm div}}$ is an integer by assumption, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ have exactly the same number of components equal to $\bar{x}$ and $\bar{x} + \delta_r$, we have

$$\mathrm{Tr}(X_i^2 - X_j^2) = 0. \tag{5.14}$$

5.4 The bound for Kullback-Leibler divergence

The following lemma is instrumental in bounding the KL-divergence.

Lemma 5.2. Denote $E_{\max} := \max_{1\leq l\leq L} \lambda_{\max}(A_l A_l^T), E_{\min} := \min_{1\leq l\leq L} \lambda_{\min}(A_l A_l^T)$. On the event $\mathcal{E}_{\rm sing}$, defined in (4.8), if $\frac{[\sigma_z^2 + x_{\max}^2\cdot E_{\max}]\cdot E_{\max}}{\left(\sigma_z^2 + x_{\min}^2 \cdot E_{\min}\right)^2} \cdot x_{\max}\delta_r < \frac{1}{4}$, we have for all $\boldsymbol{x}_i \neq \boldsymbol{x}_j \in \mathcal{S}_{\rm sep}$

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}_i} \parallel \mathbb{P}_{\boldsymbol{x}_j}) \leq 2\frac{\left(\sigma_z^2 + x_{\max}^2 E_{\max}\right)^2}{\left(\sigma_z^2 + x_{\min}^2 E_{\min}\right)^4} \sum_{l=1}^{L} \left\|A_l(X_i^2 - X_j^2)A_l^\top\right\|_{\rm HS}^2,$$

where X_i and X_j are diagonal matrices corresponding to the vectors $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{S}_{\rm sep}$.

We now apply the upper tail bound of Definition 4.10 to find a deterministic upper bound for $\sum_{l=1}^{L} \left\|A_l(X_i^2 - X_j^2)A_l^\top\right\|_{\rm HS}^2$. We set $\boldsymbol{d}_{i,j} := \boldsymbol{x}_i^2 - \boldsymbol{x}_j^2$, and define $D_{i,j} = \mathrm{diag}(\boldsymbol{d}_{i,j}))$ to get:

$$\begin{aligned}
&\|\boldsymbol{d}_{i,j}\|_\infty = \max_p |x_{i,p}^2 - x_{j,p}^2| \leq 2x_{\max}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty, \quad \|\boldsymbol{d}_{i,j}^2\|_\infty \leq 4x_{\max}^2 \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty^2, \\
&\|\boldsymbol{d}_{i,j}\|_2^2 = \sum_{p=1}^{n} \left(x_{i,p}^2 - x_{j,p}^2\right)^2 \leq 4x_{\max}^2 \sum_{p=1}^{n} \left(x_{i,p} - x_{j,p}\right)^2 = 4x_{\max}^2 \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2, \\
&\|\boldsymbol{d}_{i,j}^2\|_2^2 = \sum_{p=1}^{n} \left(x_{i,p}^2 - x_{j,p}^2\right)^4 \leq 16x_{\max}^4 \sum_{p=1}^{n} \left(x_{i,p} - x_{j,p}\right)^4 = 16x_{\max}^4 \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_4^4.
\end{aligned} \tag{5.15}$$

We choose the following values of $t_{1,i,j}$ and $t_{2,i,j}$ to apply the upper tail bound in Definition 4.10

$$t_{1,i,j} := C_{t_1}\left(x_{\max}\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|_2\sqrt{\log(mLr^2)}+x_{\max}\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|_\infty\log(mLr^2)\right);$$
$$t_{2,i,j} := C_{t_2}\log m\left(x_{\max}^2\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|_4^2\sqrt{mL(\sqrt{m}+\sqrt{n})^4\log r^2}+x_{\max}^2\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|_\infty^2(\sqrt{n}+\sqrt{m})^2\log r^2\right),$$

where C_{t_1} and C_{t_2} are two constants. To apply Definition 4.10, we note that the non-constant terms appearing in the exponent of Definition 4.10 obeys the following lower bounds

(a) $\frac{t_{1,i,j}^2}{\|\boldsymbol{d}_{i,j}\|_2^2}$ is bounded from below by $\frac{t_{1,i,j}^2}{4x_{\max}^2\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2}=\Omega(\log(mLr^2))$,

(b) $\frac{t_{1,i,j}}{\|\boldsymbol{d}_{i,j}\|_\infty}$ is bounded from below by $\frac{t_{1,i,j}}{2x_{\max}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_\infty}=\Omega(\log(mLr^2))$,

(c) $\frac{t_{2,i,j}^2}{mL(\sqrt{n}+\sqrt{m})^4\|\boldsymbol{d}_{i,j}^2\|_2^2}$ is bounded from below by $\frac{t_{2,i,j}^2}{16x_{\max}^4 mL(\sqrt{n}+\sqrt{m})^4\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_4^4}=\Omega(\log(mLr^2))$,

(d) $\frac{t_{2,i,j}}{(\sqrt{n}+\sqrt{m})^2\|\boldsymbol{d}_{i,j}\|_\infty^2}$ is bounded from below by $\Omega(\log(mLr^2))$,

(e) $\frac{t_{2,i,j}^2}{Lm^3\|\boldsymbol{d}_{i,j}\|_2^2}$ is bounded from below by $\Omega((\log m)^2\log r)$,

(f) $\frac{t_{2,i,j}}{m\|\boldsymbol{d}_{i,j}\|_\infty}$ is bounded from below by $\Omega((\log m)(\log r)))$.

In view of the above definition, consider the event

$$\mathcal{E}_{\mathrm{dcpl}} := \bigcap_{1\le i<j\le r}\left[\sum_{l=1}^{L}\|A_l(X_i^2-X_j^2)A_l^\top\|_{\mathrm{HS}}^2 < Lmt_{1,i,j}^2+Lm(m-1)\|\boldsymbol{d}_{i,j}\|_2^2+t_{2,i,j}\right]. \tag{5.16}$$

Then, using Definition 4.10, a union bound for all $1\le i<j\le r$, and $\mathrm{Tr}(D_{ij})=0, D_{ij}=\mathrm{diag}(\boldsymbol{d}_{ij}), \boldsymbol{d}_{ij}=\boldsymbol{x}_i^2-\boldsymbol{x}_j^2$ (see (5.14)), the above display implies that for sufficiently large constants C_{t_1}, C_{t_2}, we have

$$\begin{aligned}
&\mathbb{P}(\mathcal{E}_{\mathrm{dcpl}}^c\cap\widetilde{\mathcal{E}}_{\mathrm{maxsing}})\\
&\le\sum_{1\le i<j\le r}\mathbb{P}\left(\left[\sum_{l=1}^{L}\|A_lD_{ij}A_l^\top\|_{\mathrm{HS}}^2>Lm\left(\mathrm{Tr}(D_{ij})+t_{1,i,j}\right)^2+Lm(m-1)\|\boldsymbol{d}_{i,j}\|_2^2+t_{2,i,j}\right]\cap\widetilde{\mathcal{E}}_{\mathrm{maxsing}}\right)\\
&\le r^2\exp\left\{-\tilde{C}\left(\log(mLr^2)+(\log m)(\log r)\right)\right\},
\end{aligned} \tag{5.17}$$

for a large constant $\tilde{C}$. Hence, by making C_{t_1}, C_{t_2} large enough such that $\tilde{C}>10$, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{\mathrm{dcpl}}\cap\widetilde{\mathcal{E}}_{\mathrm{maxsing}}) &\ge 1-\mathbb{P}(\mathcal{E}_{\mathrm{dcpl}}^c\cap\widetilde{\mathcal{E}}_{\mathrm{maxsing}})-\mathbb{P}(\widetilde{\mathcal{E}}_{\mathrm{maxsing}}^c)\\
&\ge 1-r^2\exp\left\{-10\left(\log(mLr^2)+(\log m)(\log r)\right)\right\}-2mL\exp(-cn)\\
&\overset{(a)}{\ge} 1-\frac{1}{(rmL)^8}-2mL\exp(-cn).
\end{aligned} \tag{5.18}$$

In view of the above, on the high-probability event $\mathcal{E}_{\text{dcpl}} \cap \widetilde{\mathcal{E}}_{\text{maxsing}}$, we have for each $1 \le i < j \le r$

$$
\begin{aligned}
&\sum_{l=1}^{L} \left\| A_l (X_i^2 - X_j^2) A_l^\top \right\|_{\text{HS}}^2 \le Lmt_{1,i,j}^2 + Lm(m-1)\|\boldsymbol{d}_{i,j}\|_2^2 + t_{2,i,j} \\
&\overset{(a)}{\le} CLmx_{\max}^2 \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 \log(mLr^2) + CLm \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_\infty^2 \log^2(mLr^2) + 4Lm(m-1)x_{\max}^2 \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \\
&\quad + C \log m \left(x_{\max}^2 \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_4^2 \sqrt{mL(\sqrt{m} + \sqrt{n})^4 \log r^2} + \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_\infty^2 (\sqrt{n} + \sqrt{m})^2 \log r^2 \right) \\
&\overset{(b)}{\le} CLmx_{\max}^2 \frac{kn}{N_{\text{div}}} \delta_r^2 \log(mLr^2) + CLm\delta_r^2 \log^2(mLr^2) + 4Lm(m-1)x_{\max}^2 \frac{kn}{N_{\text{div}}} \delta_r^2 \\
&\quad + C \log m \left(x_{\max}^2 \sqrt{\frac{kn}{N_{\text{div}}}} \delta_r^2 \sqrt{mL(\sqrt{m} + \sqrt{n})^4 \log r^2} + \delta_r^2 (\sqrt{n} + \sqrt{m})^2 \log r^2 \right),
\end{aligned}
$$

where (a) followed by using the inequality $(a+b)^2 \le 2a^2 + 2b^2$, and for (b) we have used the fact that by our construction of $\mathcal{X}_k^{\text{finite}}$, $\boldsymbol{x}_i - \boldsymbol{x}_j$ is $\frac{2kn}{N_{\text{div}}}$-sparse for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{S}_{\text{sep}}$, and thus $\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_\infty \le \delta_r$, $\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 \le \frac{kn}{N_{\text{div}}} \delta_r^2$, and $\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_4^2 \le \sqrt{\frac{kn}{N_{\text{div}}}} \delta_r^2$. Hence, restricting to the event $\mathcal{E}_{\text{dcpl}} \cap \widetilde{\mathcal{E}}_{\text{maxsing}} \cap \mathcal{E}_{\text{sing}}$, together with Definition 5.2 and $E_{\max} \approx n$, we have for constant $\bar{C} := C_{x_{\min}, x_{\max}} > 0$

$$
\begin{aligned}
&\beta_r := \max_{1 \le i < j \le r} \mathrm{KL}(\mathbb{P}_i \parallel \mathbb{P}_j) \\
&\le 2 \frac{\left(\sigma_z^2 + x_{\max}^2 \cdot E_{\max}\right)^2}{\left(\sigma_z^2 + x_{\min}^2 \cdot \frac{1}{4}(\sqrt{n} - \sqrt{m})^2\right)^4} \max_{1 \le i < j \le r} \sum_{l=1}^{L} \left\| A_l (X_i^2 - X_j^2) A_l^\top \right\|_{\text{HS}}^2 \\
&\le \frac{\bar{C}}{\max(\sigma_z^4, n^2)} \Bigg(Lm \frac{kn}{N_{\text{div}}} \delta_r^2 \log(mLr^2) + Lm\delta_r^2 \log^2(mLr^2) + Lm(m-1) \frac{kn}{N_{\text{div}}} \delta_r^2 \\
&\qquad\qquad + \log m \sqrt{\frac{kn}{N_{\text{div}}}} \delta_r^2 \sqrt{mL(\sqrt{m} + \sqrt{n})^4 \log r^2} + \delta_r^2 (\sqrt{n} + \sqrt{m})^2 (\log m)(\log r^2) \Bigg) \\
&\le \frac{\bar{C}\delta_r^2 m^2 nLk}{\max(\sigma_z^4, n^2) N_{\text{div}}} \left(\frac{\log(mLr^2)}{m} + \frac{\log^2(mLr^2) N_{\text{div}}}{mnk} + 1 + \sqrt{\frac{nN_{\text{div}} \log r^2}{Lkm^3/(\log m)^2}} + \frac{(\log m)(\log r^2) N_{\text{div}}}{Lm^2} \right) \\
&\le \Theta_{x_{\min}, x_{\max}}(1) \frac{m^2 nLk}{\max(\sigma_z^4, n^2) N_{\text{div}}} \delta_r^2, \qquad (5.19)
\end{aligned}
$$

where the last inequality followed by factoring out δ_r^2 and using the following inequalities that are consequences of Definition 5.1, alongside our assumptions $\log m = \Theta(\log n)$, $\log L = O(\log n)$, and there exists $\varepsilon \in (0, 1/2)$ such that $k \le n^{1-2\varepsilon}$, $\max(\sigma_z^4, m^2, n^2) k \log n \le m^2 n^{1-\varepsilon} L$.

- $\log(mLr^2) \le \log m + \log L + 2\log r < m/3$ for all large m, L, n as $\log m = \Theta(\log n)$, $\log L = O(\log n)$ and $\log r = \Theta(\log n)$ from Definition 5.1.
- Similar to above, we have $\frac{\log^2(mLr^2) N_{\text{div}}}{mnk} \le 1$ for all large m, L, n, as $\frac{N_{\text{div}}}{nk} < n^{-(1-\epsilon)}$ from (5.10).

- $\frac{nN_{\text{div}}\log r^2}{Lkm^3/(\log m)^2} \leq 1$ as $\frac{N_{\text{div}}}{k} \leq n^{\epsilon}$ and hence, using $\max(\sigma_z^4, m^2, n^2)k\log n \leq m^2 n^{1-\varepsilon} L$ we get

$$\frac{nN_{\text{div}}\log r^2}{Lkm^3/(\log m)^2} \leq \frac{n^{1+\varepsilon}\log r^2}{Lm^3/(\log m)^2} \lesssim \frac{\max(\sigma_z^4, m^2, n^2)\log n}{n^{1-\varepsilon}Lm^{2.5}} \lesssim \frac{1}{k\sqrt{m}}.$$

- Finally $\frac{(\log m)(\log r^2)N_{\text{div}}}{Lm^2} \leq 1$ for all large m, L, k as $N_{\text{div}} = kn^{\varepsilon}$ and $m^2 L > kn^{1+\varepsilon}$ from the last property $\max(\sigma_z^4, m^2, n^2)k\log n \leq m^2 n^{1-\varepsilon} L$ as well.

In view of the above argument, and also Definition 5.2, our we choose δ_r satisfying the following.

- We want $\beta_r < c\log r$ for a small constant $c > 0$ to maximize the lower bound from Definition 4.3.
- To justify the application of Definition 5.2 we need $\frac{[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}] \cdot E_{\max}}{(\sigma_z^2 + x_{\min}^2 \cdot E_{\min})^2} \cdot x_{\max}\delta_r < \frac{1}{4}$.
- Since the maximum value the signals in $\mathcal{S}_{\text{sep}}$ can take is $\bar{x} + \delta_r$ and since $\mathcal{S}_{\text{sep}} \subset \mathcal{C}$ and the maximum value of the entries of the vectors in $\mathcal{C}$ is $x_{\max}$ we have to ensure that $\bar{x} + \delta_r < x_{\max}$.

In view of the above, consider the following choice of δ_r for a small constant $c_\delta > 0$ (possibly depending on $x_{\min}, x_{\max}$)

$$\delta_r^2 := c_\delta \Delta^{-2} \cdot \frac{\max(\sigma_z^4, n^2)\log r}{nm^2 L} \cdot \frac{N_{\text{div}}}{k} \tag{5.20}$$

where $r = |\mathcal{S}_{\text{sep}}|$, and $\Delta = \Delta(m, n, x_{\max}, x_{\min}) := \frac{[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}] \cdot E_{\max}}{(\sigma_z^2 + x_{\min}^2 \cdot E_{\min})^2} \cdot 2x_{\max}$. The first two conditions can be verified using the upper bound (5.19) and the definitions of δ_r^2. To check the final condition, i.e., $\bar{x} + \delta_r < x_{\max}$, we first note that since $E_{\max}$ and $E_{\min}$ depend on A_l, δ_r is a random number. Hence, to find a deterministic upper bound for δ_r we only consider the measurement matrices that belong to the event $\mathcal{E}_{\text{sing}}$ defined in (4.8). In view of the definition of $\mathcal{E}_{\text{sing}}$, in the above event we have $E_{\max} = \Theta(n)$ and $E_{\min} = \Theta(n)$. Hence,

$$\delta_r^2 = \Theta_{x_{\max}, x_{\min}}\left(\frac{\max(\sigma_z^4, n^2)\log r}{nm^2 L} \cdot \frac{N_{\text{div}}}{k}\right) = \Theta_{x_{\max}, x_{\min}}\left(\frac{\max(\sigma_z^4, n^2)\log r}{n^{1-\varepsilon}m^2 L}\right), \tag{5.21}$$

where the last identity followed by plugging (5.10) and the result of Lemma 5.1 By the assumptions we made in the main theorem, i.e., $\max(\sigma_z^4, m^2, n^2)k\log n \leq m^2 n^{1-\varepsilon} L$ and the fact that $\max(\sigma_z^4, m^2, n^2) = n^2$, we can see that $\left(\frac{n^{1+\epsilon}k\log(n/k)}{m^2 L}\right) < 1$. Hence, by choosing c_δ in the definition of δ_r in (5.20) small enough we can ensure that $\bar{x} + \delta_r < x_{\max}$ holds.

5.5 Concluding the proof

For sufficiently small $c_\delta > 0$, (5.19) ensures that we have $\beta_r \leq \frac{1}{10}\log r$. Therefore, by conditioning on $A_1, ..., A_L$ and restricting to the high probability event $\widetilde{\mathcal{E}}_{\text{maxsing}} \cap \mathcal{E}_{\text{dcpl}} \cap \mathcal{E}_{\text{sing}}$, we have $\frac{\beta_r}{\log r} \leq \frac{1}{10}$ and $\mathbb{P}\left(\widetilde{\mathcal{E}}_{\text{maxsing}} \cap \mathcal{E}_{\text{dcpl}} \cap \mathcal{E}_{\text{sing}}\right) > \frac{1}{2}$. In view of (5.13) and (5.20) we get that

$$\alpha_r^2 = \frac{ckn\delta_r^2}{N_{\text{div}}} = \Theta_{x_{\max}, x_{\min}}\left(\frac{\max(\sigma_z^4, n^2)}{m^2}\frac{k\log\left(N_{\text{div}}/k\right)}{L}\right). \tag{5.22}$$

As a consequence, by Definition 4.3 we have for any estimator $\widehat{\boldsymbol{x}}$, the lower bound

$$\max_{1\le i\le r}\mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_i\|^2}{n}\right]\ge\frac{\alpha_r^2}{4n}\left(1-\frac{\beta_r+\log 2}{\log r}\right)^2\mathbb{P}\left(\widetilde{\mathcal{E}}_{\mathrm{maxsing}}\cap\mathcal{E}_{\mathrm{dcpl}}\cap\mathcal{E}_{\mathrm{sing}}\right)=\Theta\left(\frac{\alpha_r^2}{n}\right)$$
$$=\Theta_{x_{\max},x_{\min}}\left(\frac{\max(\sigma_z^4,n^2)}{m^2n}\frac{k\log\left(N_{\mathrm{div}}/k\right)}{L}\right)=\Theta_{\varepsilon,x_{\max},x_{\min}}\left(\frac{\max(\sigma_z^4,n^2)k\log n}{m^2nL}\right).$$

6 Proof of the lower bound for the case $m\ge\frac{n}{4}$

This section will primarily establish the lower bound for the sub-case $m\ge 4n,\sigma_z^2=0$ given by

$$R_2(\mathcal{C}_k,m,n,0)=\inf_{\boldsymbol{\delta}}\sup_{\boldsymbol{x}\in\mathcal{C}_k}\mathbb{E}\left[\frac{\left\|\boldsymbol{\delta}(\vec{\boldsymbol{y}})-\boldsymbol{x}_o\right\|_2^2}{n}\right]=\Omega_{\varepsilon,x_{\min}x_{\max}}\left(\frac{k\log n}{nL}\right),\quad m\ge 4n. \tag{6.1}$$

Then, the lower bound for a general $\sigma_z^2\ge 0$ and $4n\ge m\ge\frac{n}{4}$ follow from Definition 4.1 and Definition 4.2, as the minimax risk is a non-increasing function of m and an increasing function of σ_z^2, the variance of the additive noise component. To see the above, we first fix $\sigma_z^2=0$. Note that, in view of Definition 4.1, as the risk is a non-increasing function of m, the last display implies

$$R_2(\mathcal{C}_k,m,n,0)\ge R_2(\mathcal{C}_k,4n,n,0)=\Omega_{\varepsilon,x_{\min},x_{\max}}\left(\frac{k\log n}{nL}\right),\quad\frac{n}{4}\le m\le 4n. \tag{6.2}$$

Hence, combining the lower bounds in (6.1) and (6.2) we get

$$R_2(\mathcal{C}_k,m,n,0)=\Omega_{x_{\max},x_{\min}}\left(\frac{k\log n}{nL}\right),\quad m\ge\frac{n}{4}. \tag{6.3}$$

To achieve a lower bound for a general $\sigma_z^2\le m$, when $m\ge\frac{n}{4}$, we first use that the miminax error is non-decreasing function in σ_z (Definition 4.2) to get $R_2(\mathcal{C}_k,m,n,0)\le R_2(\mathcal{C}_k,m,n,\sigma_z)$. Then, for $\sigma_z^2\le m$, combining (6.3) with the upper bound for $\sigma_z^2=m$ from Definition 2.7, for $m\ge\frac{n}{4}$, we get

$$C_1\frac{k\log n}{nL}\le R_2(\mathcal{C}_k,m,n,0)\le R_2(\mathcal{C}_k,m,n,\sigma_z)\le R_2(\mathcal{C}_k,m,n,\sqrt{m})\le C_2\frac{k\log n}{nL}, \tag{6.4}$$

where C_1,C_2 are constants depending on $x_{\min},x_{\max}$. Hence, the sandwich inequality implies

$$R_2(\mathcal{C}_k,m,n,\sigma_z)=\Theta_{x_{\max},x_{\min}}\left(\frac{k\log n}{nL}\right),\ \text{whenever } m\ge\frac{n}{4},\sigma_z^2\le m. \tag{6.5}$$

Next, suppose $m\ge\frac{n}{4}$ and $\sigma_z^2\ge m$. We observe that the only use of $m\le\frac{n}{4}$ in the proofs of Section 5, was to bound $E_{\min}$ from below. Notably, in the proof of Section 5, only places we used the lower bound on $E_{\min}$, are given in Definition 5.2 and (5.19). Then we note that we can repeat the entire analyses of lower bound in Section 5 to establish the lower bound for the case $m\ge\frac{n}{4}$ and $\sigma_z^2\ge m$ by replacing $\mathcal{E}_{\mathrm{sing}}$ with $\mathcal{E}_{\mathrm{maxsing}}$ and the lower bound for the singular value $\sigma_{\min}(\sigma_z^2I_n+A_lX_o^2A_l^\top)$ by σ_z^2. As a consequence, we have the lower bound

$$R_2(\mathcal{C}_k,m,n,\sigma_z)=\Omega_{x_{\max},x_{\min}}\left(\frac{\sigma_z^4}{m^2n}\cdot\frac{k\log n}{L}\right),\ \text{whenever } m\ge\frac{n}{4},\sigma_z^2\ge m. \tag{6.6}$$

Combining (6.6) and (6.5) for the case $m \geq \frac{n}{4}$ yields the desired minimax lower bound

$$R_2(\mathcal{C}_k, m, n, \sigma_z) = \Omega_{x_{\max}, x_{\min}}\left(\frac{\max(\sigma_z^4, m^2)}{m^2 n} \cdot \frac{k \log n}{L}\right), \quad m \geq \frac{n}{4}, \sigma_z \geq 0. \tag{6.7}$$

We complete our proof by establishing the lower bound in (6.1). We bring forward ideas related to sufficient statistics for our proof, using the following definition used throughout this section.

Definition 6.1. (Casella and Berger, 2024, Definition 6.2.1) A statistics $\boldsymbol{T}(\vec{\boldsymbol{y}})$ is sufficient for $\boldsymbol{x}_o$ if the conditional distribution of the sample $\vec{\boldsymbol{y}}$ given the value $\boldsymbol{T}(\vec{\boldsymbol{y}})$ does not depend on $\boldsymbol{x}_o$.

Define $\boldsymbol{A} = \operatorname{diag}(A_1, \ldots, A_L) \in \mathbb{R}^{mL \times nL}$ and note that $\boldsymbol{A}^\top \boldsymbol{A} = \operatorname{diag}(A_1^\top A_1, \ldots, A_L^\top A_L)$. Throughout the section we analyze the expected loss on the high probability event $\mathcal{E}'_{\text{sing}}$ defined in (4.10), where all the matrices $\{A_l^\top A_l\}_{l=1}^L$ are invertible. Then we have the following result.

Proposition 6.2. Consider the case $\sigma_z = 0$ and that the event $\mathcal{E}'_{\text{sing}}$ holds. Then $\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}}) = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \vec{\boldsymbol{y}}$ is a sufficient statistic for the parameter $\boldsymbol{x}_o$.

Proof. Note that, $\boldsymbol{T}_A(\vec{\boldsymbol{y}})$ is an one to one transformation of $\vec{\boldsymbol{y}}$ whenever $\boldsymbol{A}^\top \boldsymbol{A}$ is invertible, as $\boldsymbol{A}\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}}) = \vec{\boldsymbol{y}}$. As any one to one transformation of a sufficient statistics is also a sufficient statistics, and $\vec{\boldsymbol{y}}$ is itself a sufficient statistics, we get that $\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}})$ is a sufficient statistics. $\square$

We will use the Rao-Blackwell theorem to first bound the desired minimax risk from below using the squared error loss for the sufficient statistic $\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}})$.

Theorem 6.3 (Rao–Blackwell theorem, MSE version)**.** (Shao, 2008, Theorem 2.5) Let $\boldsymbol{\delta}(\vec{\boldsymbol{y}})$ be any estimator of the parameter $\boldsymbol{x}_o$ and $\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}})$ is a sufficent statistics for $\boldsymbol{x}_o$. Then $g(\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}})) = \mathbb{E}[\boldsymbol{\delta}(\vec{\boldsymbol{y}}) \mid \boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}})]$ is also an estimator for $\boldsymbol{x}_o$ and it provides an improved error guarantee

$$\mathbb{E}\left[\left\|\boldsymbol{\delta}(\vec{\boldsymbol{y}}) - \boldsymbol{x}_o\right\|_2^2 |\mathcal{E}'_{\text{sing}}\right] \geq \mathbb{E}\left[\left\|g(\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}})) - \boldsymbol{x}_o\right\|_2^2 |\mathcal{E}'_{\text{sing}}\right].$$

In view of the above result, we establish a lower bound to $\mathbb{E}\left[\left\|g(\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}})) - \boldsymbol{x}_o\right\|_2^2\right]$ to complete our analysis. This is provided in the result below.

Lemma 6.4. Consider the model (1.2) with $\sigma_z = 0, m \geq 4n$. Then, there exists a constant $C \geq 0$, we have

$$\inf_g \sup_{\boldsymbol{x} \in \mathcal{C}_k} \mathbb{E}\left[\left\|g(\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}})) - \boldsymbol{x}_o\right\|_2^2 |\mathcal{E}'_{\text{sing}}\right] \geq C \frac{k \log n}{L}.$$

Proof of Definition 6.4. By our construction, we have that $\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}}) = [\boldsymbol{u}_1^\top, ..., \boldsymbol{u}_L^\top]^\top$, where $\boldsymbol{u}_l = (A_l^\top A_l)^{-1} A_l \boldsymbol{y}_l \in \mathbb{R}^n$. In the above optimization problem, as the estimators we consider are all of the form $g(\boldsymbol{T}_{\boldsymbol{A}}(\vec{\boldsymbol{y}}))$, we may treat $\boldsymbol{u}_1, ..., \boldsymbol{u}_L$ as our observations and the problem transformed into recovering $\boldsymbol{x}_o \in \mathcal{C}_k$ from the simplified model

$$\boldsymbol{u}_l = X_o \boldsymbol{w}_l, \quad \boldsymbol{w}_l \sim N(0, I_n), l = 1, 2, ..., L.$$

Hence, the new data distribution for which we will apply Lemma 4.3 to derive the lower bound is

$$\mathbb{P}_{\boldsymbol{x}} \sim \otimes_{l=1}^L N(\boldsymbol{0}, \Sigma_l^{-1}(\boldsymbol{x})) = N(\boldsymbol{0}, \boldsymbol{\Sigma}^{-1}(\boldsymbol{x})),$$
$$\boldsymbol{\Sigma}(\boldsymbol{x}) := \operatorname{diag}\left(\Sigma_1(\boldsymbol{x}), ..., \Sigma_L(\boldsymbol{x})\right), \quad \Sigma_l = \Sigma_l(\boldsymbol{x}) := (\sigma_z^2 I_n + X^2)^{-1}.$$

To obtain the lower bound, we shall use Definition 4.3 and a similar discretization and construction of α_r-separated set $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_r\}$ as in Section 5.3, with $N_{\text{div}} = kn^{\varepsilon}$ and the following choice for δ_r^2, α_r^2

$$\delta_r^2 := c_\delta \frac{\log r}{nL} \cdot \frac{N_{\text{div}}}{k}, \text{ and } \frac{\alpha_r^2}{n} := \frac{k}{N_{\text{div}}} \delta_r^2, \tag{6.8}$$

for a small constant $c_\delta > 0$. Denote $\mathbb{P}_{\boldsymbol{x}_i}$ as $\mathbb{P}_i$ for $i \in \{1, \ldots, r\}$. Using the fact that by our earlier construction of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r\}$, we have $\frac{kn}{N_{\text{div}}}$ and $\boldsymbol{x}_i - \boldsymbol{x}_j$ is $\frac{2kn}{N_{\text{div}}}$-sparse for any $\boldsymbol{x}_i \neq \boldsymbol{x}_j$, which implies $\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 \leq \frac{2kn}{N_{\text{div}}} \delta_r^2$. Hence, using Definition 5.2, with $A_l = I_n$, and $E_{\max} = E_{\min} = 1$), we get

$$\begin{aligned} \beta_r &:= \max_{1 \leq i < j \leq r} \mathrm{KL}(\mathbb{P}_i \parallel \mathbb{P}_j) \leq L \frac{x_{\max}^4}{x_{\min}^8} \max_{1 \leq i < j \leq r} \left\| X_i^2 - X_j^2 \right\|_{\mathrm{HS}}^2 \\ &\leq \frac{4x_{\max}^6}{x_{\min}^8} L \max_{1 \leq i < j \leq r} \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 \leq \frac{4x_{\max}^6}{x_{\min}^8} \frac{Lkn\delta_r^2}{N_{\text{div}}} \leq \frac{1}{10} \log r, \end{aligned} \tag{6.9}$$

for sufficiently small c_δ in (6.8) to guarantee a small δ_r^2. Now it follows from Definition 4.3 that

$$\inf_{\boldsymbol{\delta}} \sup_{\boldsymbol{x} \in \mathcal{C}_k} \mathbb{E}\left[\frac{\left\|g(\boldsymbol{T_A}(\vec{\boldsymbol{y}})) - \boldsymbol{x}_o\right\|_2^2}{n} \middle| \mathcal{E}'_{\text{sing}} \right] \geq \frac{\alpha_r^2}{4n} \left(1 - \frac{\beta_r + \log 2}{\log r}\right)^2 \mathbb{P}[\mathcal{E}'_{\text{sing}}] = \Theta_{\varepsilon, x_{\min}, x_{\max}} \left(\frac{k \log n}{nL}\right).$$

where the last equality followed using (6.8), as $N_{\text{div}} = kn^{\varepsilon}$ and $\frac{\alpha_r^2}{n} = \frac{k\delta_r^2}{N_{\text{div}}} = \Theta_{x_{\min}, x_{\max}} \left(\frac{k \log\left(\frac{N_{\text{div}}}{k}\right)}{nL}\right)$. $\square$

In view of the last display, it follows from Theorem 6.3 and Definition 6.4 to conclude (6.1)

$$\inf_{\boldsymbol{\delta}} \sup_{\boldsymbol{x} \in \mathcal{C}_k} \mathbb{E}\left[\frac{\left\|\boldsymbol{\delta}(\vec{\boldsymbol{y}}) - \boldsymbol{x}_o\right\|_2^2}{n} \right] \geq \inf_{\boldsymbol{\delta}} \sup_{\boldsymbol{x} \in \mathcal{C}_k} \mathbb{E}\left[\frac{\left\|g(\boldsymbol{T_A}(\vec{\boldsymbol{y}})) - \boldsymbol{x}_o\right\|_2^2}{n} \middle| \mathcal{E}'_{\text{sing}} \right] \mathbb{P}\left(\mathcal{E}'_{\text{sing}}\right) = \Omega\left(\frac{k \log n}{nL}\right).$$

7 Proof of the upper bound

7.1 General strategy

Note that without loss of generality, we may assume $a = x_{\max} - x_{\min}$ and $b = 1$.

We provide the proof of all the technical results in this section later in Appendix E. We will show that the desired upper bound is achieved by the maximum likelihood estimator

$$\widehat{\boldsymbol{x}}_o = \arg\min_{\boldsymbol{x} \in \mathcal{C}} f(\boldsymbol{x}), \tag{7.1}$$

where $\mathcal{C}$ is the class of all possible signals, and the negative log-likelihood $f(\boldsymbol{x})$ is defined as

$$f(\boldsymbol{x}) = \sum_{l=1}^{L} \log\det\left(\sigma_z^2 I_m + A_l X^2 A_l^\top\right) + \sum_{l=1}^{L} \boldsymbol{y}_l^\top \left(\sigma_z^2 I_m + A_l X^2 A_l^\top\right)^{-1} \boldsymbol{y}_l \tag{7.2}$$

For the entirety of the analysis in this section, we will restrict ourselves to the following event

$$\mathcal{S} = \left\{ \{A_l\}_{i=1}^L : \sigma_{\min}(A_l A_l^\top) \geq E_{\min}, \quad \sigma_{\max}(A_l A_l^\top) \leq E_{\max}, \quad l = 1, \ldots, L \right\}. \tag{7.3}$$

This is the only place where we impose restrictions on the singular values of $\{A_l\}_{l=1}^L$. To get to the specific minimax risk guarantees in different regimes we will choose appropriate values of $E_{\max}, E_{\min}$. In particular, we have the following considerations.

- Case I ($n \geq 4m$): We will choose $E_{\max} = \frac{9}{4}(\sqrt{n} + \sqrt{m})^2$ and $E_{\min} = \frac{1}{4}(\sqrt{n} - \sqrt{m})^2$. In that case the event $\mathcal{S}$ satisfies $\mathcal{E}_{\text{sing}} \subseteq \mathcal{S}$, where $\mathcal{E}_{\text{sing}}$ is given as in (4.8). This implies $\mathbb{P}[\mathcal{S}] \geq \mathbb{P}[\mathcal{E}_{\text{sing}}] \geq 1 - Le^{-cn}$ for some constant $c > 0$.
- Case II ($n < 4m$): We will choose $E_{\max} = \frac{9}{4}(\sqrt{n} + \sqrt{m})^2$ and $E_{\min} = 0$. In that case the event $\mathcal{S}$ satisfies $\mathcal{E}_{\text{maxsing}} \subseteq \mathcal{S}$, where $\mathcal{E}_{\text{maxsing}}$ is given as in (4.6). This implies $\mathbb{P}[\mathcal{S}] \geq \mathbb{P}[\mathcal{E}_{\text{maxsing}}] \geq 1 - Le^{-cn}$ for some constant $c > 0$.

In other words we also have

$$\mathbb{P}[\mathcal{S}] \geq 1 - Le^{-cn}, \quad \text{for all } L, m, n. \tag{7.4}$$

Consider the following notations for simplifying the presentation. Let $\{\Sigma_l\}_{l=1}^L$ be the collection of inverses of the covariance matrix $\mathbb{E}\left[\boldsymbol{y}_l \boldsymbol{y}_l^\top | A_l\right]$ given by

$$\Sigma_l = \Sigma_l(\boldsymbol{x}) := (\sigma_z^2 I_m + A_l X^2 A_l^\top)^{-1}, \quad l = 1, \ldots, L. \tag{7.5}$$

Define the vector $\vec{\boldsymbol{y}} \in \mathbb{R}^{mL}$ and block-diagonal matrix $\boldsymbol{\Sigma}(\boldsymbol{x}) \in \mathbb{R}^{mL \times mL}$ as the collection of all the observations and the inverse covariance matrices over different looks

$$\vec{\boldsymbol{y}}^\top := (\boldsymbol{y}_1^\top, ..., \boldsymbol{y}_L^\top), \quad \boldsymbol{\Sigma}(\boldsymbol{x}) := \operatorname{diag}\left(\Sigma_1(\boldsymbol{x}), ..., \Sigma_L(\boldsymbol{x})\right), \quad \boldsymbol{\Sigma}_o = \boldsymbol{\Sigma}(\boldsymbol{x}_o), \quad \widehat{\boldsymbol{\Sigma}}_o = \boldsymbol{\Sigma}(\hat{\boldsymbol{x}}_o). \tag{7.6}$$

In view of the above notations, we can rewrite the negative log-likelihood in (7.2) as

$$f(\boldsymbol{x}) = -\log\det(\boldsymbol{\Sigma}(\boldsymbol{x})) + \vec{\boldsymbol{y}}^\top \boldsymbol{\Sigma}(\boldsymbol{x}) \vec{\boldsymbol{y}} \tag{7.7}$$

Now we proceed with the proof. Our proof strategy draws inspiration from the empirical loss minimization literature, such as Fan and Gu (2024), Fan et al. (2025), to achieve a parametric error rate in the sample size L by comparing the negative log-likelihood for the estimator $\widehat{\boldsymbol{x}}_o$ and the true parameter $\boldsymbol{x}_o$, that also turns out to be the minimax rate. Since $\widehat{\boldsymbol{x}}_o$ is the minimizer of (7.1), we have

$$f(\widehat{\boldsymbol{x}}_o) \leq f(\boldsymbol{x}_o). \tag{7.8}$$

For a fixed $\boldsymbol{x}$, define $\overline{f}(\boldsymbol{x})$ as the function of conditional expectation of $f(\boldsymbol{x})$ given $A_1, \ldots, A_L$

$$\overline{f}(\boldsymbol{x}) := \mathbb{E}[f(\boldsymbol{x}) \mid A_1, ..., A_L] = -\log\det\boldsymbol{\Sigma}(\boldsymbol{x}) + \operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{x})\boldsymbol{\Sigma}(\boldsymbol{x}_o)^{-1}\right). \tag{7.9}$$

Simplifying the expression for $f(\widehat{\boldsymbol{x}}_o) - f(\boldsymbol{x}_o)$, with the notations in (7.6) we get

$$
\begin{aligned}
& f(\widehat{\boldsymbol{x}}_o) - f(\boldsymbol{x}_o) \\
&= \vec{\boldsymbol{y}}^\top \left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right) \vec{\boldsymbol{y}} - \mathrm{Tr}\left[\boldsymbol{\Sigma}_o^{-1}\left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\right] + \mathrm{Tr}\left[\boldsymbol{\Sigma}_o^{-1}\left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\right] - \log\det(\widehat{\boldsymbol{\Sigma}}_o) + \log\det(\boldsymbol{\Sigma}_o) \\
&= \vec{\boldsymbol{y}}^\top \left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right) \vec{\boldsymbol{y}} - \mathrm{Tr}\left[\boldsymbol{\Sigma}_o^{-1}\left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\right] + \overline{f}(\hat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o).
\end{aligned} \tag{7.10}
$$

Therefore, in view of (7.8) we get

$$
\vec{\boldsymbol{y}}^\top \left(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o\right) \vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)) \geq \overline{f}(\hat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o). \tag{7.11}
$$

Our following approach is to find an upper bound for the left side in terms of $\|\widehat{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2$ and a lower bound for the right side in terms of $\|\widehat{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2$, and simplify the inequality to get an upper bound for $\|\widehat{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2$. Throughout the rest of the draft we will use the following notation

$$
C_{n,m,\sigma_z} := C(n, m, \sigma_z, x_{\max}, x_{\min}) = c\frac{\sigma_z^2 + x_{\max}^2 E_{\max}}{\sigma_z^2 + x_{\min}^2 E_{\min}}, \tag{7.12}
$$

where $c > 0$ is a large universal constant. Note that in the regime $n \geq 4m$, C_{n,m,σ_z} is of constant order as long as $x_{\max}, x_{\min}$ are of constant order. We will use similar expressions similar to C_{n,m,σ_z} in the analysis of the case $n < 4m$, the related details will be presented later according to requirements.

Establishing an upper bound on $\vec{\boldsymbol{y}}^\top(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o))$: The main challenge in the analysis is the dependency of $\widehat{\boldsymbol{\Sigma}}_o$ on $\vec{\boldsymbol{y}}$, which prevents us from directly applying concentration inequalities to bound $\vec{\boldsymbol{y}}^\top(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}}$. To resolve this issue, we use a δ-net argument, as will be clarified below. Consider a δ-net of the set $\mathcal{C}_k$, denoted by $N_{\delta_{\rm net}}(\mathcal{C}_k)$, with the choice of δ to be discussed later. Define $\widetilde{\boldsymbol{x}}_o$ as the closest vector in $N_{\delta_{\rm net}}(\mathcal{C}_k)$ to $\boldsymbol{x}_o$, i.e.,

$$
\widetilde{\boldsymbol{x}}_o = \mathrm{argmin}_{\boldsymbol{x} \in N_{\delta_{\rm net}}(\mathcal{C}_k)} \|\widehat{\boldsymbol{x}}_o - \boldsymbol{x}\|_2. \tag{7.13}
$$

We will use the following notations for the rest of the section

$$
\widetilde{\boldsymbol{\Sigma}}_o = \boldsymbol{\Sigma}(\widetilde{\boldsymbol{x}}_o), \quad \widetilde{X}_o = \mathrm{diag}(\widetilde{\boldsymbol{x}}_o), \quad \widetilde{\boldsymbol{\Sigma}}_o = \boldsymbol{\Sigma}(\widetilde{\boldsymbol{x}}), \quad \widetilde{X} = \mathrm{diag}(\widetilde{\boldsymbol{x}}), \quad \widetilde{\boldsymbol{x}} \in N_{\delta_{\rm net}}(\mathcal{C}_k). \tag{7.14}
$$

Then in view of triangle inequality we get

$$
\begin{aligned}
& \left|\vec{\boldsymbol{y}}^\top(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o))\right| \\
& \leq \left|\vec{\boldsymbol{y}}^\top(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o))\right| + \left|\vec{\boldsymbol{y}}^\top(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o))\right|.
\end{aligned} \tag{7.15}
$$

We use an union bound argument to control the first term above, uniformly over all possible choices of $\widetilde{\boldsymbol{x}} \in N_{\delta_{\rm net}}(\mathcal{C}_k)$. This is done in the following result.

Lemma 7.1. There exist constants $c_1, c_2, c_3, c_4 > 0$ such that the following holds with probability $1 - Le^{-cn} - e^{-c_1 Lk \log((x_{\max} - x_{\min})n/\delta_{\rm net})}$

$$
\left|\vec{\boldsymbol{y}}^\top(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o))\right| \leq b_1\sqrt{\mathscr{X}} + b_1', \quad \text{for all } \widetilde{\boldsymbol{x}}_o \in N_{\delta_{\rm net}}(\mathcal{C}_k),
$$

where, with the notation in (7.12), $b_1, b_1', \mathscr{L}$ are defined as

$$b_1 = c_3 \sqrt{k \log\left(\frac{(x_{\max} - x_{\min})n}{\delta_{\text{net}}}\right)}, \quad b_1' = c_4 C_{n,m,\sigma_z} k \log\left(\frac{(x_{\max} - x_{\min})n}{\delta_{\text{net}}}\right) \cdot x_{\max}^2, \\ \mathscr{L} = \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)). \tag{7.16}$$

The following result controls the final term of (7.15).

Lemma 7.2. Let C_{n,m,σ_z} be as in (7.12) and denote $b_2 = (C_{n,m,\sigma_z})^2 mL\delta_{\text{net}}$. There exist constants $c_1, c_2 > 0$ such that the following holds with probability $1 - Le^{-c_1 n} - e^{-c_2 mL}$

$$\left| \vec{\boldsymbol{y}}^\top (\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o)) \right| \leq b_2.$$

Combining Lemma 7.2 with Lemma 7.1, in view of (7.15) we have

$$\left| \vec{\boldsymbol{y}}^\top (\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)) \right| \leq b_1\sqrt{\mathscr{L}} + b_1' + b_2 \tag{7.17}$$

Establishing a lower bound on $\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o)$: To find the lower bound, we use the decomposition

$$\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o) = \overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\widetilde{\boldsymbol{x}}_o) + \overline{f}(\widetilde{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o), \tag{7.18}$$

with $\widetilde{\boldsymbol{x}}_o$ as in (7.13). The first term, $\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\widetilde{\boldsymbol{x}}_o)$ can be bounded by $C_{n,m,\sigma_z} x_{\max} n\delta_{\text{net}}$ using the fact that $\tilde{\boldsymbol{x}}_o$ is chosen to be at most δ_{net} distance away from $\hat{\boldsymbol{x}}_o$. We bound $\overline{f}(\widetilde{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o)$ using the following result.

Lemma 7.3. Assume that $\sigma_z^2 I_m + A_l \widetilde{X}_o^2 A_l^\top$ and $\sigma_z^2 I_m + A_l X_o^2 A_l^\top, 1 \leq l \leq L$, are invertible. Then,

$$\overline{f}(\widetilde{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o) \geq \frac{1}{2(1+\widetilde{\lambda}_{\max})^2} \mathrm{Tr}\left(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\right), \tag{7.19}$$

where $\widetilde{\lambda}_{\max} > 0$ is the maximum singular value of $\boldsymbol{\Sigma}_o^{-\frac{1}{2}}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\boldsymbol{\Sigma}_o^{-\frac{1}{2}}$. Moreover, $\widetilde{\lambda}_{\max} \leq C_{n,m,\sigma_z}$ on the event $\mathcal{S}$ in (7.3).

The following result controls $|\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\widetilde{\boldsymbol{x}}_o)|$ for a given δ_{net}.

Lemma 7.4. $|\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\widetilde{\boldsymbol{x}}_o)| \leq C_{n,m,\sigma_z} \cdot x_{\max} n\delta_{\text{net}} \ll 1$ with probability $1 - Le^{-cn}$ for some $c > 0$.

Combining the above results, in view of (7.18) we have, with probability $1 - L\exp(-cn)$,

$$\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o) \geq \frac{\mathscr{L}}{(C_{n,m,\sigma_z})^2} - 1. \tag{7.20}$$

Simplifying the quadratic inequality: Combining (7.20) and (7.17), in view of (7.11), we have

$$\mathbb{P}\left[\frac{\mathscr{L}}{(C_{n,m,\sigma_z})^2} \leq b_1\sqrt{\mathscr{L}} + b_1' + b_2 + 1\right] \geq 1 - e^{-c_1 Lk\log((x_{\max}-x_{\min})n/\delta_{\text{net}})} - Le^{-cn} - \exp\left(-cmL\right). \tag{7.21}$$

Rewrite the last inequality as $az^2 - bz - c \le 0$, with $z = \sqrt{\mathscr{X}}$, $a = \frac{1}{(C_{n,m,\sigma_z})^2}$, $b = b_1$, $c = b_1' + b_2 + 1$. As $z = \sqrt{\mathscr{X}} > 0$, z^2 is smaller than the square of the positive root of $az^2 - bz - c = 0$, which implies

$$\mathscr{X} = z^2 \le \left(\frac{-b + \sqrt{b^2 + 4ac}}{2a} \right)^2 \le \left(\frac{-b + \sqrt{b^2 + 4ac}}{2a} \right) \left(\frac{b + \sqrt{b^2 + 4ac}}{2a} \right) = \frac{c}{a}, \tag{7.22}$$

where the second inequality followed as $a, b, c > 0$. Using the notations from, (7.12), Lemma 7.1 and Lemma 7.2 we get

$$\mathscr{X} = \frac{b_1' + b_2}{a} = (C_{n,m,\sigma_z})^2 \left(c_3 k \log \left(\frac{(x_{\max} - x_{\min})n}{\delta_{\text{net}}} \right) \cdot x_{\max}^2 + (C_{n,m,\sigma_z})^2 mL\delta_{\text{net}} \right).$$

Choose $\delta_{\text{net}} = \frac{x_{\max}}{n^5}$ and recall $mL \le n^4 k \log n$ from Definition 2.7. Then, from the last display we use (7.21) to get for a constant $C > 0$

$$\mathbb{P} \left[\mathscr{X} \le C \cdot (C_{n,m,\sigma_z})^2 k \log n \right] = 1 - O \left(n^{-ckL} + L \exp(-cn) + 2 \exp(-cmL) \right). \tag{7.23}$$

Finding a lowerbound for $\mathscr{X}$: In view of Definition 4.14, using the block structure of $\mathbf{\Sigma}_o, \widetilde{\mathbf{\Sigma}}_o$ given in (7.14), we have on the event $\mathcal{S}$,

$$\begin{aligned} \mathscr{X} &= \operatorname{Tr} \left[\mathbf{\Sigma}_o^{-1} \left(\widetilde{\mathbf{\Sigma}}_o - \mathbf{\Sigma}_o \right) \mathbf{\Sigma}_o^{-1} \left(\widetilde{\mathbf{\Sigma}}_o - \mathbf{\Sigma}_o \right) \right] \\ &= \sum_{l=1}^{L} \operatorname{Tr} \left[(\Sigma_l(\boldsymbol{x}_o)^{-1} (\Sigma_l(\widetilde{\boldsymbol{x}}_o) - \Sigma_l(\boldsymbol{x}_o)) \Sigma_l(\boldsymbol{x}_o)^{-1} (\Sigma_l(\widetilde{\boldsymbol{x}}_o) - \Sigma_l(\boldsymbol{x}_o)) \right] \\ &\ge \frac{1}{(C_{n,m,\sigma_z})^2 \left(\sigma_z^2 + x_{\max}^2 E_{\max} \right)^2} \sum_{l=1}^{L} \left\| A_l (\widehat{X}_o^2 - X_o^2) A_l^\top \right\|_{\text{HS}}^2, \end{aligned} \tag{7.24}$$

where C_{n,m,σ_z} is as in (7.12). The lower bound on $\mathscr{X}$ is completed with the following lower bound on $\sum_{l=1}^{L} \left\| A_l (\widehat{X}_o^2 - X_o^2) A_l^\top \right\|_{\text{HS}}^2$. The proof follows from Lemma 4.10 and is given in Subsection E.5.

Lemma 7.5. The following holds true with a probability $1 - \exp(-2k \log n) - mL \exp(-cn)$

$$\begin{aligned} &\sum_{l=1}^{L} \| A_l (\widetilde{X}_o^2 - X_o^2) A_l^\top \|_{\text{HS}}^2 \\ &\ge 4m(m-1)L x_{\min}^2 \|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2 - 4C x_{\max}^2 \|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2 \log m \sqrt{mL} n \sqrt{k \log n} - C x_{\max}^4 nk \log m \log n. \end{aligned}$$

Final upper bound on $\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2$: We combine (7.24), (7.23), and Lemma 7.5 to summarize the above in terms of the following quadratic inequality with respect to $\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2$, that holds with a probability $1 - O \left(n^{-ckL} + L \exp(-cn) + \exp(-cmL) \right)$

$$\begin{gathered} a \|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2 - b \|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2 - d \le 0 \\ a = \frac{C_1 m(m-1) L x_{\min}^2}{(C_{n,m,\sigma_z})^2 (\sigma_z^2 + x_{\max}^2 E_{\max})^2}, \quad b = \frac{C_2 x_{\max}^2 n \log m \sqrt{mLk \log n}}{(C_{n,m,\sigma_z})^2 (\sigma_z^2 + x_{\max}^2 E_{\max})^2}, \\ d = \frac{C_3 x_{\max}^4 nk \log m \log n}{C \cdot (C_{n,m,\sigma_z})^2 (\sigma_z^2 + x_{\max}^2 E_{\max})^2} + C \cdot (C_{n,m,\sigma_z})^2 k \log n. \end{gathered} \tag{7.25}$$

In view of an argument similar to (7.22) we have with a probability $1-O\left(n^{-ckL}+L\exp(-cn)+\exp\left(-cmL\right)\right)$

$$\frac{1}{n}\|\widetilde{\boldsymbol{x}}_o-\boldsymbol{x}_o\|_2^2 \le \frac{d}{na} \le \frac{C_3 x_{\max}^4}{x_{\min}^2}\frac{k\log m\log n}{m^2L}+\frac{(C_{n,m,\sigma_z})^4(\sigma_z^2+x_{\max}^2E_{\max})^2k\log n}{nm^2L}. \tag{7.26}$$

This implies, in view of $\frac{1}{n}\|\widetilde{\boldsymbol{x}}_o-\boldsymbol{x}_o\|_2^2\le x_{\max}^2$,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\|\widetilde{\boldsymbol{x}}_o-\boldsymbol{x}_o\|_2^2\right] \le{}& \frac{C_3 x_{\max}^4}{x_{\min}^2}\frac{k\log m\log n}{m^2L}+\frac{(C_{n,m,\sigma_z})^4(\sigma_z^2+x_{\max}^2E_{\max})^2k\log n}{nm^2L}\\ &+C_1x_{\max}^2(n^{-ckL}+L\exp(-cn)+\exp\left(-cmL\right)).\end{aligned} \tag{7.27}$$

As $\|\widetilde{\boldsymbol{x}}_o-\widehat{\boldsymbol{x}}_o\|_2\le\delta_{\rm net}\le\frac{x_{\max}}{n^5}$ from the definition in (7.13), we continue the last display to get

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\|\widehat{\boldsymbol{x}}_o-\boldsymbol{x}_o\|_2^2\right] \le 2C_4\Bigg\{&\frac{x_{\max}^4}{x_{\min}^2}\frac{k\log m\log n}{m^2L}+\frac{(C_{n,m,\sigma_z})^4(\sigma_z^2+x_{\max}^2E_{\max})^2k\log n}{nm^2L}\\ &+x_{\max}^2(n^{-ckL}+L\exp(-cn)+\exp\left(-cmL\right))+\frac{x_{\max}^2}{n^{10}}\Bigg\}.\end{aligned} \tag{7.28}$$

Note by our assumption $\log m\ll n$. Therefore the first term has a slower growth rate compared to the second term.

7.2 Proof of Theorem 2.7

We first consider the subcase $n\ge 4m$, and the subcase $n<4m$ with $\sigma_z^2\ge m$. Then we have C_{n,m,σ_z} is of a constant order and $E_{\max}=\Theta(m+n)$. In view of (7.28), the above implies

$$\mathbb{E}\left[\frac{1}{n}\|\widehat{\boldsymbol{x}}_o-\boldsymbol{x}_o\|_2^2\right]\le C_{x_{\max},x_{\min}}\left\{\frac{\max(\sigma_z^4,m^2,n^2)k\log n}{m^2nL}+n^{-ckL}+L\exp(-cn)+\exp\left(-cmL\right)+\frac{1}{n^{10}}\right\},$$

for a constant $C>0$ depending on $x_{\min},x_{\max}$. We now focus on the remaining scenario of $n<4m$ with $\sigma_z^2<m$. Here, using the fact that the error is a non-decreasing function of σ_z (Definition 4.2), we obtain the upper bound

$$R_2(\mathcal{C}_k,m,n,\sigma_z)\le R_2(\mathcal{C}_k,m,n,\sqrt{m})=C\left\{\frac{k\log n}{nL}+n^{-ckL}+L\exp(-cn)+\exp\left(-cmL\right)+\frac{1}{n^{10}}\right\},$$

for a constant $C>0$ depending on $x_{\min},x_{\max}$. The above coincides with our desired upper bound, completing the result.

Acknowledgment

Arian Maleki is supported in part by NSF grant DMS-2515716, and by ONR award N00014-23-1-2371.

References

Achim, A., Bezerianos, A., and Tsakalides, P. (2001). Novel bayesian multiscale method for speckle removal in medical ultrasound images. *IEEE Transactions on Medical Imaging*, 20(8):772–783.

Argenti, F., Lapini, A., Bianchi, T., and Alparone, L. (2013). A tutorial on speckle reduction in synthetic aperture radar images. *Geoscience and Remote Sensing Magazine, IEEE*, 1:6–35.

Arias-Castro, E., Salmon, J., and Willett, R. (2012). Oracle inequalities and minimax rates for nonlocal means and related adaptive kernel-based methods. *SIAM Journal on Imaging Sciences*, 5(3):944–992.

Bakhshizadeh, M., Maleki, A., and Jalali, S. (2020). Using black-box compression algorithms for phase retrieval. *IEEE Transactions on Information Theory*, 66(12):7978–8001.

Bank, D., Koenigstein, N., and Giryes, R. (2023). Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374.

Bate, T., O'Keefe, D., Spencer, M. F., and Pellizzari, C. J. (2022). Experimental validation of model-based digital holographic imaging using multi-shot data. In *Unconventional Imaging and Adaptive Optics 2022*, volume 12239, pages 83–94. SPIE.

Bellec, P. C., Lecué, G., and Tsybakov, A. B. (2018). Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642.

Bianco, V., Memmolo, P., Leo, M., Montrésor, S., Distante, C., Paturzo, M., Picart, P., Javidi, B., and Ferraro, P. (2018). Strategies for reducing speckle noise in digital holography. *Light: Science & Applications*, 7:1–16.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732.

Cai, T. T., Li, X., and Ma, Z. (2016). Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics*, 44(5):2221 – 2251.

Candes, E. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215.

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Candès, E. J. and Su, W. (2015). Slope is adaptive to unknown sparsity and asymptotically minimax. *CoRR*.

Casella, G. and Berger, R. (2024). *Statistical inference*. Chapman and Hall/CRC.

Chen, X., Hou, Z., Metzler, C. A., Maleki, A., and Jalali, S. (2024). Bagged deep image prior for recovering images in the presence of speckle noise. *arXiv preprint arXiv:2402.15635*.

Chen, X., Jana, S., Metzler, C. A., Maleki, A., and Jalali, S. (2025). Multilook coherent imaging: Theoretical guarantees and algorithms. *arXiv preprint arXiv:2505.23594*.

Chen, Y. and Candès, E. J. (2017). Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on pure and applied mathematics*, 70(5):822–883.

Chen, Y., Chi, Y., Fan, J., and Ma, C. (2019). Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37.

Dasari, K., Anjaneyulu, L., Jayasri, P., and Prasad, A. (2015). Importance of speckle filtering in image classification of sar data. In *2015 International Conference on Microwave, Optical and Communication Engineering (ICMOCE)*, pages 349–352.

Davidson, K. and Szarek, S. (2001). *Local operator theory, random matrices and Banach spaces*, page 317–366. North-Holland, Amsterdam.

De la Pena, V. and Giné, E. (2012). *Decoupling: from dependence to independence*. Springer Science & Business Media.

De Vries, F. (1998). Speckle reduction in sar imagery by various multi-look techniques. Technical report, TNO Physics and Electronics Laboratory.

DeVore, R., Nowak, R. D., Parhi, R., Petrova, G., and Siegel, J. W. (2025). Optimal recovery meets minimax estimation.

Donoho, D. L. (1999). Wedgelets: Nearly minimax estimation of edges. *the Annals of Statistics*, 27(3):859–897.

Donoho, D. L. and Johnstone, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455.

Donoho, D. L. and Johnstone, I. M. (1994b). Minimax risk over lp-balls for lp-error. *Probability theory and related fields*, 99(2):277–303.

Donoho, D. L. and Johnstone, I. M. (1998a). Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3):879–921.

Donoho, D. L. and Johnstone, I. M. (1998b). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879 – 921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):301–337.

Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

Duchi, J. (2007). Derivations for linear algebra and optimization. *Berkeley, California*, 3(1):2325–5870.

Fan, J. and Gu, Y. (2024). Factor augmented sparse throughput deep relu neural networks for high dimensional regression. *Journal of the American Statistical Association*, 119(548):2680–2694.

Fan, J., Jana, S., Kulkarni, S., and Yin, Q. (2025). Factor informed double deep learning for average treatment effect estimation. *arXiv preprint arXiv:2508.17136*.

Ghosh, S., Guo, Y., Weng, H., and Maleki, A. (2025). Signal-to-noise ratio aware minimax analysis of sparse linear regression. *arXiv preprint arXiv:2501.13323*.

Graham, R. L., Knuth, D. E., and Patashnik, O. (1994). *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, MA, 2nd edition.

Guo, Y., Ghosh, S., Weng, H., and Maleki, A. (2024). A note on the minimax risk of sparse linear regression. *arXiv preprint arXiv:2405.05344*.

Hand, P., Leong, O., and Voroninski, V. (2018). Phase retrieval under a generative prior. *Advances in Neural Information Processing Systems*, 31.

Hanson, D. and Wright, F. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *Ann Math Statist*, 42.

Heckel, R. and Hand, P. (2018). Deep decoder: Concise image representations from untrained non-convolutional networks. *arXiv preprint arXiv:1810.03982*.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press.

Jalali, S. and Maleki, A. (2016). From compression to compressed sensing. *Applied and Computational Harmonic Analysis*, 40(2):352–385.

Kerkyacharian, G. and Picard, D. (1992). Density estimation in besov spaces. *Statistics and Probablility Letters*, 13:15–24.

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction, Lecture Notes in Statistics*, volume 82. Springer-Verlag.

Lopez-Martinez, C. and Fabregas, X. (2003). Polarimetric sar speckle noise model. *IEEE Transactions on Geoscience and Remote Sensing*, 41(10):2232–2242.

Ma, J., Xu, J., and Maleki, A. (2019). Optimization-based amp for phase retrieval: The impact of initialization and ℓ_2 regularization. *IEEE Transactions on Information Theory*, 65(6):3600–3629.

Maleki, A., Narayan, M., and Baraniuk, R. G. (2012). Suboptimality of nonlocal means for images with sharp edges. *Applied and Computational Harmonic Analysis*, 33(3):370–387.

Maleki, A., Narayan, M., and Baraniuk, R. G. (2013). Anisotropic nonlocal means denoising. *Applied and Computational Harmonic Analysis*, 35(3):452–482.

Malekian, R., Xing, H., and Maleki, A. (2025). Is speckle noise more challenging to mitigate than additive noise? *IEEE Transactions on Information Theory*.

Metzler, C. A., Maleki, A., and Baraniuk, R. G. (2016). From denoising to compressed sensing. *IEEE Transactions on Information Theory*, 62(9):5117–5144.

Nemirovskii, A. (1985). Nonparametric estimation of smooth regression func tions. *Soviet J. of Computer and Systems Sciences*, 23:1–11.

Nemirovskii A.S., P. B. and Tsybakov, A. (1985). Rate of convergence of nonparametric estimators of maximum-likelihood type. *Problems of Infor mation Transmission*, 21:258–272.

Ortega-Martinez, A., Zimmermann, B., Cheng, X., Li, X., Yucel, M. A., and Boas, D. A. (2019). Contribution of speckle noise in near-infrared spectroscopy measurements. *Journal of Biomedical Optics*, 24(10):105003–105003.

Racine, R., Walker, G. A., Nadeau, D., Doyon, R., and Marois, C. (1999). Speckle noise and the detection of faint companions. *Publications of the Astronomical Society of the Pacific*, 111(759):587.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994.

Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268.

Schmitt, J., Xiang, S., and Yung, K. (1999). Speckle in optical coherence tomography. *Journal of Biomedical Optics*.

Shao, J. (2008). *Mathematical Statistics*. Springer Science & Business Media.

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473.

Su, W., Bogdan, M., and Candes, E. (2017). False discoveries occur early on the lasso path. *The Annals of statistics*, pages 2133–2150.

Tsybakov, A. (1986). Robust reconstruction of functions by a local-approximation method. *Problems of Information Transmission*, 22:133–146.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6(none):38 – 90.

Weng, H., Zheng, L., Maleki, A., and Wang, X. (2016). Phase transition and noise sensitivity of ℓ_p-minimization for $0 \leq p \leq 1$. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 675–679. IEEE.

Yu, B. (1997). *Assouad, Fano, and Le Cam*, pages 423–435. Springer New York, New York, NY.

Zhang, H., Chi, Y., and Liang, Y. (2016). Provable non-convex phase retrieval with outliers: Median truncatedwirtinger flow. In *International conference on machine learning*, pages 1022–1031. PMLR.

Zhang, H., Liang, Y., and Chi, Y. (2017). A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 18(141):1–35.

Zhou, W., Jalali, S., and Maleki, A. (2022). Compressed sensing in the presence of speckle noise. *IEEE Transactions on Information Theory*, 68(10):6964–6980.

Zhou, W., Jalali, S., and Maleki, A. (2024). Corrections to "compressed sensing in the presence of speckle noise". *IEEE Transactions on Information Theory*, 70(12):9352–9354.

A Proofs of Examples of Section 2.1

Proof of Example 2.2

Note that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^k$ we have $\|g(\boldsymbol{x}) - g(\boldsymbol{y})\|_2 \leq M\|\boldsymbol{x} - \boldsymbol{y}\|_2$. If $B_1, ..., B_r$ are the balls in a ball of radius $\sqrt{k}$ centered at 0 containing ε-covering of $[0,1]^k$, then $g(\mathcal{C})$ is contained in a ball of radius $M\sqrt{k}$. Hence, using Definition 4.6, we have $N_\varepsilon\left(g([0,1]^k)\right) \leq \left(\frac{2M\sqrt{k}}{\varepsilon} + 1\right)^k$.

Proof of Example 2.3

We have

$$\mathcal{S}_k = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_0 \leq k\} = \cup_{1 \leq i_1 \leq \cdots \leq i_k \leq n} \left\{\boldsymbol{x} \in \mathbb{R}^n : x_j = 0, j \neq i_1, \cdots, i_k\right\}. \tag{A.1}$$

Hence, $\mathcal{S}_k$ is a union of $\binom{n}{k}$ k-dimensional subspaces $\left\{\boldsymbol{x} \in \mathbb{R}^n : x_j = 0, j \neq i_1, \cdots, i_k\right\}$. According to by Definition 4.6 the intersection of each of these subspaces and $B_2(1)$ can be covered by at most $\left(\frac{2}{\varepsilon} + 1\right)^k$ balls of radius ε. Hence, $N_\varepsilon(\mathcal{C}) \leq \binom{n}{k}\left(\frac{2}{\varepsilon} + 1\right)^k$. To obtain the lower bound we notice that according to Definition 4.6, in order to cover one of the subspaces we need $\left(\frac{1}{\epsilon}\right)^k$. The proof of the $\kappa(\mathcal{C}) = k$ is straightforward and is hence skipped.

Proof of Definition 2.4

$f(\theta) = D\theta$ is a $\sigma_{\max}(D)$-Lipchitz function of θ. Hence, combining Definition 2.3 with a proof similar to the one presented for Definition 2.2 establishes the result.

Proof of Definition 2.5

Note $\mathcal{C} \subset D^{-1}(\mathcal{S}_k \cap B_2(0,1))$. By direct calculation,

$$D^{-1} := \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ 0 & 0 & \ddots & 1 \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

Hence $\sigma_{\max}(D^{-1}) \leq \left\|D^{-1}\right\|_{\mathrm{HS}} = \sqrt{\frac{n(n+1)}{2}} < n$. So this is a special case of Definition 2.4.

Proof of definition 2.6

We showed $\sigma_{\min}(D) \geq \frac{1}{n}$ in the proof of Definition 2.5. Hence, we have

$$\sigma_{\min}(D^{M+1}) \geq (\sigma_{\min}(D))^{M+1} \geq \left(\frac{1}{n}\right)^{M+1}.$$

B The extension of Definition 2.8 to general $a, b > 0$

Suppose our signal class $\mathcal{C} \in \mathcal{F}_{a,b,k,n}$ satisfies the polynomial complexity $N_\varepsilon(\mathcal{C}) \le \left(\frac{an^b}{\varepsilon}\right)^k$ for general $a, b > 0$. Note that we can make the transform

$$\left(\frac{an^b}{\varepsilon}\right)^k = \left(\frac{(x_{\max} - x_{\min})n}{\varepsilon'}\right)^{k'}$$

where $k' := bk$ and $\varepsilon' := \frac{(x_{\max}-x_{\min})\varepsilon^{1/b}}{a^{1/b}}$. Put $C' = \frac{\varepsilon'}{\varepsilon}\mathcal{C} \subset [x'_{\min}, x'_{\max}]^n$ where $x'_{\min} = \frac{\varepsilon'}{\varepsilon}x_{\min}$ and $x'_{\max} = \frac{\varepsilon'}{\varepsilon}x_{\max}$, and we have

$$N_{\varepsilon'}(\mathcal{C}') = N_\varepsilon(\mathcal{C}) \le \left(\frac{an^b}{\varepsilon}\right)^k = \left(\frac{(x_{\max} - x_{\min})n}{\varepsilon'}\right)^{k'}. \tag{B.1}$$

This means $\mathcal{C} \in \mathcal{F}_{a,b,k,n}$ if and only if $\mathcal{C}' \in \frac{\varepsilon'}{\varepsilon}\mathcal{F}_{a_0,b_0,k',n}$ where $a_0 = x_{\max} - x_{\min}$ and $b_0 = 1$. Hence by Definition 2.8, we have when $\varepsilon' \in (0, 1/2)$, $k' \le n^{1-2\varepsilon'}$, and $\max(\sigma_z^4, m^2, n^2)k\log n \le m^2 n^{1-\varepsilon'} L$

$$\begin{aligned}
&\sup_{\mathcal{C}\in\mathcal{F}_{a,b,k,n}} R_2(\mathcal{C}', m, n, \sigma_z) \\
=&\sup_{\mathcal{C}'\in\frac{\varepsilon'}{\varepsilon}\mathcal{F}_{a_0,b_0,k',n}} R_2(\mathcal{C}', m, n, \sigma_z) \\
=&\Omega_{\varepsilon', x'_{\max}, x'_{\min}}\left(\frac{\max(\sigma_z^4, m^2, n^2)k\log n}{m^2 nL}\right) \\
=&\Omega_{\varepsilon, x_{\max}, x_{\min}, a, b}\left(\frac{\max(\sigma_z^4, m^2, n^2)k\log n}{m^2 nL}\right).
\end{aligned} \tag{B.2}$$

C Proof of auxiliary lemmas from Section 4

C.1 Proof of Lemma 4.14

In the following steps $\otimes$ denotes the Kronecker product. We have

$$\begin{aligned}
&\mathrm{Tr}\left[\Sigma^{-1}(\widetilde{\Sigma} - \Sigma)\Sigma^{-1}(\widetilde{\Sigma} - \Sigma))\right] \\
=&\,\mathrm{Vec}(\widetilde{\Sigma} - \Sigma)^\top \left[\Sigma^{-1} \otimes \Sigma^{-1}\right] \mathrm{Vec}(\widetilde{\Sigma} - \Sigma) \\
\ge&\left\|\mathrm{Vec}(\widetilde{\Sigma} - \Sigma)\right\|_2^2 \lambda_{\min}(\Sigma^{-1} \otimes \Sigma^{-1}) \\
=&\|\widetilde{\Sigma} - \Sigma\|_{\mathrm{HS}}^2 \lambda_{\min}^2(\Sigma^{-1}) = \|\widetilde{\Sigma} - \Sigma\|_{\mathrm{HS}}^2 \lambda_{\min}^2(\sigma_z^2 I_m + A_l X_o^2 A_l^\top) \\
=&\|\widetilde{\Sigma} - \Sigma\|_{\mathrm{HS}}^2 \left[\sigma_z^2 + \lambda_{\min}(A_l X_o^2 A_l^\top)\right]^2 \\
\ge&\|\widetilde{\Sigma} - \Sigma\|_{\mathrm{HS}}^2 \left(\sigma_z^2 + x_{\min}^2 \lambda_{\min}(A_l A_l^\top)\right)^2.
\end{aligned}$$

On the other hand, using $\widetilde{\Sigma} - \Sigma = \Sigma(\Sigma^{-1} - \widetilde{\Sigma}^{-1})\widetilde{\Sigma} = \Sigma A(X^2 - \widetilde{X}^2)A^T\widetilde{\Sigma}$, we have

$$\begin{aligned}\|\widetilde{\Sigma} - \Sigma\|_{\mathrm{HS}} \geq& \lambda_{\min}(\Sigma)\lambda_{\min}(\widetilde{\Sigma}) \left\|A(\widetilde{X}_o^2 - X_o^2)A^\top\right\|_{\mathrm{HS}} \\ \geq& \frac{\left\|A(\widetilde{X}^2 - X^2)A^\top\right\|_{\mathrm{HS}}}{\lambda_{\max}(\sigma_z^2 I_m + AX^2A^\top)\lambda_{\max}(\sigma_z^2 I_m + A\widetilde{X}_o^2A^\top)} \geq \frac{\left\|A(\widetilde{X}^2 - X^2)A^\top\right\|_{\mathrm{HS}}}{\left(\sigma_z^2 + x_{\max}^2\lambda_{\max}(AA^\top)\right)^2}.\end{aligned}$$

This proves the lower bound. The proof of the upper bound is similar. Note that

$$\mathrm{Tr}\left[\Sigma^{-1}(\widetilde{\Sigma} - \Sigma)\Sigma^{-1}(\widetilde{\Sigma} - \Sigma))\right] \leq \|\widetilde{\Sigma} - \Sigma\|_{\mathrm{HS}}^2 \left(\sigma_z^2 + x_{\max}^2\lambda_{\max}(A_lA_l^\top)\right)^2, \tag{C.1}$$

and

$$\|\widetilde{\Sigma} - \Sigma\|_{\mathrm{HS}} \leq \frac{\left\|A(\widetilde{X}^2 - X^2)A^\top\right\|_{\mathrm{HS}}}{\lambda_{\min}(\sigma_z^2 I_m + AX^2A^\top)\lambda_{\min}(\sigma_z^2 I_m + A\widetilde{X}_o^2A^\top)} \leq \frac{\left\|A(\widetilde{X}^2 - X^2)A^\top\right\|_{\mathrm{HS}}}{\left(\sigma_z^2 + x_{\min}^2\lambda_{\min}(AA^\top)\right)^2}.$$

C.2 Proof of Definition 4.12

Define $B = A^TA$. Then, we have

$$\|AD\|_{\mathrm{HS}}^2 = \mathrm{Tr}(DA^TAD) = \mathrm{Tr}(DBD) = \sum_i D_{ii}^2 B_{ii}. \tag{C.2}$$

Note that if $\mathbf{e}_i$ is the unit vector with a one in the i^{th} position and zeros elsewhere. , then

$$|B_{ii}| = e_i^T B e_i \leq \sigma_{\max}(B) = \sigma_{\max}^2(A) \tag{C.3}$$

Combining (C.2) and (C.3) establishes the desired result.

C.3 Proof of Definition 4.10

The proof closely follows that of (Zhou et al., 2024, Lemma 4 and 5). However, we obtain sharper results with revised techniques, which we present here. For $\{A_l\}_{l=1}^L \in \mathbb{R}^{m\times n}, D \in \mathbb{R}^{L\times L}$, define

$$\boldsymbol{A} = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & A_L \end{bmatrix} \in \mathbb{R}^{mL\times nL}, \quad \mathbf{D} = \begin{bmatrix} D & 0 & \cdots & 0 \\ 0 & D & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & D \end{bmatrix} \in \mathbb{R}^{nL\times nL}. \tag{C.4}$$

For $1 \leq l \leq L$, let $\mathbf{a}_{l,i}^\top$ denote the i^{th} row of matrix A_l. We have

$$\|\mathbf{D}\boldsymbol{A}\|_{\mathrm{HS}}^2 = \sum_{l=1}^L \|A_lDA_l^\top\|_{\mathrm{HS}}^2 = \sum_{l=1}^L\sum_{i=1}^m\sum_{j=1}^m |\mathbf{a}_{l,i}^\top D\mathbf{a}_{l,j}|^2 = \sum_{l=1}^L\sum_{i\neq j} |\mathbf{a}_{l,i}^\top D\mathbf{a}_{l,j}|^2 + \sum_{l=1}^L\sum_{i=1}^m |\mathbf{a}_{l,i}^\top D\mathbf{a}_{l,i}|^2. \tag{C.5}$$

First note that by using the union bound and Definition 4.8, we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{l=1}^{L}\sum_{i=1}^{m}|\mathbf{a}_{l,i}^{\top}D\mathbf{a}_{l,i}|^{2} > Lm\left(\mathrm{Tr}(D)+t_1\right)^{2}\right)\\
\leq & Lm\mathbb{P}\left(|\mathbf{a}_{l,i}^{\top}D\mathbf{a}_{l,i}|^{2} > \left(\mathrm{Tr}(D)+t_1\right)^{2}\right)\\
= & Lm\mathbb{P}\left(|\mathbf{a}_{l,i}^{\top}D\mathbf{a}_{l,i}| > \mathrm{Tr}(D)+t_1\right) \leq 2mL\exp\left(-c\min\left(\frac{t_1^2}{K^4\|\boldsymbol{d}\|_2^2},\frac{t_1}{K^2\|\boldsymbol{d}\|_\infty}\right)\right),
\end{aligned} \tag{C.6}$$

where K the subgaussian norm of each element of $\mathbf{a}_{l,i}$ and $K\in[1,2]$ is a fixed number. For the off-diagonal part of (C.5), first note that

$$\mathbb{E}\left[\sum_{l=1}^{L}\sum_{i\neq j}|\mathbf{a}_{l,i}^{\top}D\mathbf{a}_{l,j}|^{2}\right] = Lm(m-1)\sum_{i=1}^{n}d_i^2 = Lm(m-1)\|\mathbf{d}\|_{\mathbf{2}}^{\mathbf{2}}. \tag{C.7}$$

By Definition 4.7, there exists a constant $C>0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{l=1}^{L}\sum_{i\neq j}|\mathbf{a}_{l,i}^{\top}D\mathbf{a}_{l,j}|^{2} - Lm(m-1)\sum_{i=1}^{n}d_i^2\right| > t\right)\\
\leq & C\mathbb{P}\left(C\left|\sum_{l=1}^{L}\sum_{i\neq j}|\mathbf{a}_{l,i}^{\top}D\tilde{\mathbf{a}}_{l,j}|^{2} - Lm(m-1)\sum_{i=1}^{n}d_i^2\right| > t\right)\\
= & C\mathbb{P}\left(C\left|\sum_{l=1}^{L}\sum_{i=1}^{m}\mathbf{a}_{l,i}^{\top}D\left(\sum_{i\neq j=1}^{m}\tilde{\mathbf{a}}_{l,j}\tilde{\mathbf{a}}_{l,j}^{\top}\right)D\mathbf{a}_{l,i} - Lm(m-1)\sum_{i=1}^{n}d_i^2\right| > t\right),
\end{aligned} \tag{C.8}$$

where $\tilde{\mathbf{a}}_{l,j}$'s denote the independent copies of $\mathbf{a}_{l,j}$'s for $1\leq l\leq L$ and $1\leq j\leq m$. Define $\tilde{A}_l$ as the $m\times n$ Gaussian matrix whose rows are $\tilde{\mathbf{a}}_{l,1}^{\top},\ldots,\tilde{\mathbf{a}}_{l,m}^{\top}$, i.e.,

$$\widetilde{A}_l := \begin{bmatrix}\tilde{\mathbf{a}}_{l,1}^{\top}\\ \vdots \\ \tilde{\mathbf{a}}_{l,m}^{\top}\end{bmatrix} \tag{C.9}$$

Also, let $\tilde{A}_{l,\setminus i}$ denote the matrix that is constructed by removing the i^{th} row of $\tilde{A}_l$. Define

$$\mathbf{F} := \begin{bmatrix} F_1 & 0 & \cdots & 0\\ 0 & F_2 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & F_L\end{bmatrix}$$

where

$$F_l := \begin{bmatrix} D\tilde{A}_{l,\backslash 1}^\top \tilde{A}_{l,\backslash 1} D & 0 & \dots & 0 \\ 0 & D\tilde{A}_{l,\backslash 2}^\top \tilde{A}_{l,\backslash 2} D & \dots & 0 \\ 0 & 0 & \dots & D\tilde{A}_{l,\backslash m}^\top \tilde{A}_{l,\backslash m} D \end{bmatrix}.$$

and

$$\vec{\boldsymbol{v}} := [\boldsymbol{v}_1^\top, ..., \boldsymbol{v}_L^\top], \quad \boldsymbol{v}_l := [\mathbf{a}_{l,1}^\top, \dots, \mathbf{a}_{l,m}^\top].$$

Let $\mathbb{P}_{\tilde{\mathbf{A}}}(\cdot) := \mathbb{P}\left[\cdot \mid \tilde{A}_1, ..., \tilde{A}_L\right]$. Note that for any event $\mathcal{E}$, by definition $\mathbb{P}_{\tilde{\mathbf{A}}}(\mathcal{E}) = \mathbb{E}[\mathbf{1}_{\mathcal{E}} \mid \tilde{\mathbf{A}}]$. It follows that

$$\begin{aligned} &\mathbb{P}\left(C\left|\sum_{l=1}^{L}\sum_{i=1}^{m}\mathbf{a}_{l,i}^\top D\left(\sum_{i\neq j=1}^{m}\tilde{\mathbf{a}}_{l,j}\tilde{\mathbf{a}}_{l,j}^\top\right)D\mathbf{a}_{l,i} - Lm(m-1)\sum_{i=1}^{n}d_i^2\right| > t_2\right) \\ =&\mathbb{E}\left[\mathbb{P}_{\tilde{\mathbf{A}}}\left(C\left|\vec{\boldsymbol{v}}^\top\mathbf{F}\vec{\boldsymbol{v}} - \mathbb{E}\left[\vec{\boldsymbol{v}}^\top\mathbf{F}\vec{\boldsymbol{v}}|\tilde{\mathbf{A}}\right]\right| \geq t_2/2\right)\right] + \mathbb{P}\left(C\left|\mathbb{E}\left[\vec{\boldsymbol{v}}^\top\mathbf{F}\vec{\boldsymbol{v}}|\tilde{\mathbf{A}}\right] - Lm(m-1)\sum_{i=1}^{n}d_i^2\right| \geq t_2/2\right). \end{aligned} \tag{C.10}$$

Note that from Definition 4.8 (with fixed $\tilde{\mathbf{A}}$) we have

$$\mathbb{P}_{\tilde{\mathbf{A}}}\left(C\left|\vec{\boldsymbol{v}}^\top\mathbf{F}\vec{\boldsymbol{v}} - \mathbb{E}\left[\vec{\boldsymbol{v}}^\top\mathbf{F}\vec{\boldsymbol{v}}|\tilde{\mathbf{A}}\right]\right| \geq t/2\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{4C^2K^4\|\mathbf{F}\|_{\mathrm{HS}}^2}, \frac{t}{2CK^2\|\mathbf{F}\|_2}\right)\right). \tag{C.11}$$

Define the event $\widetilde{\mathcal{E}}_{\mathrm{maxsing}} := \bigcap_{l=1}^{L}\bigcap_{i=1}^{m}\left\{\sigma_{\max}(\tilde{A}_{l,\backslash i}) \leq \frac{3}{2}(\sqrt{m}+\sqrt{n})\right\}$, and define $E_{\max} = \frac{9}{4}(\sqrt{m}+\sqrt{n})^2$. By Definition 4.9, we have that $\mathbb{P}(\mathcal{E}_{\mathrm{maxsing}}) \geq 1 - mL\exp(-cn)$. Restricted to the event $\widetilde{\mathcal{E}}_{\mathrm{maxsing}}$, we have

$$\|\mathbf{F}\|_2 = \max_{1\leq l\leq L}\max_{1\leq i\leq m}\sigma_{\max}\left(D\tilde{A}_{l,\backslash i}^\top\tilde{A}_{l,\backslash i}D\right) \leq E_{\max}\|\boldsymbol{d}\|_\infty^2. \tag{C.12}$$

and

$$\begin{aligned} &\|\mathbf{F}\|_{\mathrm{HS}}^2 = \sum_{l=1}^{L}\|F_l\|_{\mathrm{HS}}^2 \\ =&\sum_{l=1}^{L}\sum_{i=1}^{m}\left\|D\tilde{A}_{l,\backslash i}^\top\tilde{A}_{l,\backslash i}D\right\|_{\mathrm{HS}}^2 = \sum_{l=1}^{L}\sum_{i=1}^{m}\mathrm{Tr}\left[D\tilde{A}_{l,\backslash i}^\top\tilde{A}_{l,\backslash i}DD\tilde{A}_{l,\backslash i}^\top\tilde{A}_{l,\backslash i}D\right] \\ =&\sum_{l=1}^{L}\sum_{i=1}^{m}\mathrm{Tr}\left[\tilde{A}_{l,\backslash i}^\top\tilde{A}_{l,\backslash i}DD\tilde{A}_{l,\backslash i}^\top\tilde{A}_{l,\backslash i}DD\right] \\ \overset{(a)}{\leq}&\sum_{l=1}^{L}\sum_{i=1}^{m}(\sum_{p}\sigma_p(\tilde{A}_{l,\backslash i}^\top\tilde{A}_{l,\backslash i}DD))^2 = \sum_{l=1}^{L}\sum_{i=1}^{m}\left\|\tilde{A}_{l,\backslash i}^\top\tilde{A}_{l,\backslash i}DD\right\|_{\mathrm{HS}}^2 \leq Lm\left[E_{\max}\right]^2\|\boldsymbol{d}^2\|_2^2, \end{aligned} \tag{C.13}$$

where to obtain inequality (a) we have used (4.13) and to obtain the last inequality we have used Lemma 4.12. Combining, (C.11), (C.12), and (C.13) we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{P}_{\tilde{\mathbf{A}}}\left(C\left|\vec{\boldsymbol{v}}^\top \mathbf{F}\vec{\boldsymbol{v}} - \mathbb{E}\left[\vec{\boldsymbol{v}}^\top \mathbf{F}\vec{\boldsymbol{v}}|\tilde{\mathbf{A}}\right]\right| \geq t_2/2\right)\right] \\
\leq & 2\exp\left(-c\min\left(\frac{4t_2^2}{81C^2K^4\|\boldsymbol{d}^2\|_2^2 mL(\sqrt{n}+\sqrt{m})^4}, \frac{2t_2}{9CK^2\|\boldsymbol{d}\|_\infty^2(\sqrt{n}+\sqrt{m})^2}\right)\right).
\end{aligned} \tag{C.14}
$$

Now note that

$$
\begin{aligned}
&\mathbb{P}\left(C\left|\mathbb{E}\left[\vec{\boldsymbol{v}}^\top \mathbf{F}\vec{\boldsymbol{v}}|\tilde{\mathbf{A}}\right] - Lm(m-1)\sum_{i=1}^n d_i^2\right| \geq t/2\right) \\
=&\mathbb{P}\left(C\left|\left[\sum_{l=1}^L\sum_{i=1}^m\sum_{j\neq i, j=1}^m \tilde{\boldsymbol{a}}_{l,j}^\top D^2\tilde{\boldsymbol{a}}_{l,j}\right] - Lm(m-1)\sum_{i=1}^n d_i^2\right| \geq t/2\right) \\
\overset{(a)}{\leq}& m\mathbb{P}\left(\left|\sum_{l=1}^L\sum_{j\neq i_0, j=1}^m \tilde{\boldsymbol{a}}_{l,j}^\top D^2\tilde{\boldsymbol{a}}_{l,j} - L(m-1)\sum_{i=1}^n d_i^2\right| \geq t/2m\right) \\
\overset{(b)}{\leq}& 2m\exp\left(-c\min\left(\frac{t^2}{4K^4m^2\left\|\mathbf{D}^2_{(m-1)L}\right\|_{\mathrm{HS}}^2}, \frac{t}{2K^2m\left\|\mathbf{D}^2_{(m-1)L}\right\|_2}\right)\right) \\
\leq& 2m\exp\left(-c\min\left(\frac{t^2}{4K^4Lm^3\|\boldsymbol{d}^2\|_2^2}, \frac{t}{2K^2m\|\boldsymbol{d}^2\|_\infty}\right)\right),
\end{aligned} \tag{C.15}
$$

where to obtain inequality (a) we have used the union bound (the distribution for $\sum_{j\neq i_0, j=1}^m \tilde{\boldsymbol{a}}_{l,j}^\top D^2\tilde{\boldsymbol{a}}_{l,j}$ is the same for all $1\leq i_0\leq m$), and to obtain (b) we have used Definition 4.8 for the $L(m-1)n\times L(m-1)n$ matrix

$$
\mathbf{D}^2_{(m-1)L} := \begin{bmatrix} D^2 & 0 & \cdots & 0 \\ 0 & D^2 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & D^2 \end{bmatrix}
$$

Finally, combining (C.6),(C.8), (C.10), (C.14), and (C.15) we have that

$$
\begin{aligned}
&\mathbb{P}\left(\left[\sum_{l=1}^{L}\|A_l D A_l^\top\|_{\mathrm{HS}}^2 > Lm\left(\operatorname{Tr}(D)+t_1\right)^2 + Lm(m-1)\|\boldsymbol{d}\|_2^2 + t_2\right] \cap \widetilde{\mathcal{E}}_{\mathrm{maxsing}}\right)\\
\leq&\mathbb{P}\left(\sum_{l=1}^{L}\sum_{i=1}^{m}|\mathbf{a}_{l,i}^\top D\mathbf{a}_{l,i}|^2 > Lm\left(\operatorname{Tr}(D)+t_1\right)^2\right)\\
&+\mathbb{P}\left(\left[\left|\sum_{l=1}^{L}\sum_{i\neq j}|\mathbf{a}_{l,i}^\top D\mathbf{a}_{l,j}|^2 - Lm(m-1)\sum_{i=1}^{n} d_i^2\right| > t_2\right] \cap \widetilde{\mathcal{E}}_{\mathrm{maxsing}}\right)\\
\leq&2mL\exp\left(-c\min\left(\frac{t_1^2}{K^4\|\boldsymbol{d}\|_2^2}, \frac{t_1}{K^2\|\boldsymbol{d}\|_\infty}\right)\right)\\
&+2C\exp\left(-c\min\left(\frac{4t_2^2}{81C^2K^4\|\boldsymbol{d}^2\|_2^2 mL(\sqrt{n}+\sqrt{m})^4}, \frac{2t_2}{9CK^2\|\boldsymbol{d}\|_\infty^2(\sqrt{n}+\sqrt{m})^2}\right)\right)\\
&+2m\exp\left(-c\min\left(\frac{t_2^2}{4K^4Lm^3\|\boldsymbol{d}^2\|_2^2}, \frac{t_2}{2K^2Lm^2\|\boldsymbol{d}^2\|_\infty}\right)\right),
\end{aligned}
$$

for some constants c and C. On the other hand,

$$
\begin{aligned}
&\mathbb{P}\left(\left[\sum_{l=1}^{L}\|A_l D A_l^\top\|_{\mathrm{HS}}^2 < Lm(m-1)\|\boldsymbol{d}\|_2^2 - t\right] \cap \widetilde{\mathcal{E}}_{\mathrm{maxsing}}\right)\\
\leq&\mathbb{P}\left(\left[\sum_{l=1}^{L}\sum_{i\neq j}|\mathbf{a}_{l,i}^\top D\mathbf{a}_{l,j}|^2 - Lm(m-1)\sum_{i=1}^{n} d_i^2 < -t\right] \cap \widetilde{\mathcal{E}}_{\mathrm{maxsing}}\right)\\
\leq&\mathbb{P}\left(\left[\left|\sum_{l=1}^{L}\sum_{i\neq j}|\mathbf{a}_{l,i}^\top D\mathbf{a}_{l,j}|^2 - Lm(m-1)\sum_{i=1}^{n} d_i^2\right| > t\right] \cap \widetilde{\mathcal{E}}_{\mathrm{maxsing}}\right)\\
\leq&2C\exp\left(-c\min\left(\frac{4t^2}{81C^2K^4\|\boldsymbol{d}^2\|_2^2 mL(\sqrt{n}+\sqrt{m})^4}, \frac{2t}{9CK^2\|\boldsymbol{d}\|_\infty^2(\sqrt{n}+\sqrt{m})^2}\right)\right)\\
&+2m\exp\left(-c\min\left(\frac{t^2}{4K^4Lm^3\|\boldsymbol{d}^2\|_2^2}, \frac{t}{2K^2Lm^2\|\boldsymbol{d}^2\|_\infty}\right)\right),
\end{aligned}
$$

for some constants c and C.

C.4 Proof of Definition 4.1 and Definition 4.2

Proof of Definition 4.1. Let $m < m'$ be two positive integers and consider two scenarios of our speckle noise model:

$$\begin{aligned} \boldsymbol{y}_l =& A_l X_o \boldsymbol{w}_l + \boldsymbol{z}_l, \text{ for } l = 1, \dots, L. \\ \boldsymbol{y}'_l =& A'_l X_o \boldsymbol{w}'_l + \boldsymbol{z}'_l, \text{ for } l = 1, \dots, L. \end{aligned} \tag{C.16}$$

where we have independent $\boldsymbol{w}_l \sim \mathcal{N}(0, I_n)$, $\boldsymbol{w}'_l \sim \mathcal{N}(0, I_n)$, $\boldsymbol{z}_l \sim \mathcal{N}(0, \sigma_z I_m)$, and $\boldsymbol{z}'_l \sim \mathcal{N}(0, \sigma_z I_{m'})$. We would like to show

$$R_2(\mathcal{C}_k, m', n, \sigma_z) \leq R_2(\mathcal{C}_k, m, n, \sigma_z). \tag{C.17}$$

Indeed, for each $1 \leq l \leq L$, we look at each component of vectors $\boldsymbol{y}_l$ and $\boldsymbol{y}'_l$: For $1 \leq i \leq m$ and $1 \leq i' \leq m'$,

$$y_{l,i} = \sum_j A_{l,ij} x_{o,j} w_{l,j} + z_{l,i}, \tag{C.18}$$

$$y'_{l,i'} = \sum_j A'_{l,i'j} x_{o,j} w'_{l,j} + z'_{l,i'}. \tag{C.19}$$

For $1 \leq l \leq L$, let $\boldsymbol{y}'_l|_m$ denote the truncation of the m'-dimensional vector on its first m components.

Let $\mathscr{E}_m$ denote the collection of all estimators for the first scenario, let $\mathscr{E}'_{m'}$ denote the collection of all estimators for the second scenario, and let $\mathscr{E}'_m \subset \mathscr{E}'_{m'}$ denote the subcollection of estimators for the second scenario that only use the information of $\boldsymbol{y}'_l|_m$ for $1 \leq l \leq L$.

In these two scenarios, we construct estimators $\widehat{\boldsymbol{x}}(\boldsymbol{y}_1, ..., \boldsymbol{y}_L)$ and $\widehat{\boldsymbol{x}}'(\boldsymbol{y}'_1, ..., \boldsymbol{y}'_L)$, and we can always view $\widehat{\boldsymbol{x}}(\boldsymbol{y}_1, ..., \boldsymbol{y}_L)$ as a special case of $\widehat{\boldsymbol{x}}'(\boldsymbol{y}'_1, ..., \boldsymbol{y}'_L)$ where we only use the information of the truncations $\boldsymbol{y}'_1|_m, ..., \boldsymbol{y}'_L|_m$. Therefore

$$\begin{aligned} & \inf_{\widehat{\boldsymbol{x}} \in \mathscr{E}_m} \sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\widehat{\boldsymbol{x}}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_L) - \boldsymbol{x}_o\|_2^2\right] \\ =& \inf_{\widehat{\boldsymbol{x}}' \in \mathscr{E}'_m} \sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\widehat{\boldsymbol{x}}'(\boldsymbol{y}'_1|_m, \dots, \boldsymbol{y}'_L|_m) - \boldsymbol{x}_o\|_2^2\right] \\ \geq& \inf_{\widehat{\boldsymbol{x}}' \in \mathscr{E}'_{m'}} \sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\widehat{\boldsymbol{x}}'(\boldsymbol{y}'_1, \dots, \boldsymbol{y}'_L) - \boldsymbol{x}_o\|_2^2\right], \end{aligned} \tag{C.20}$$

and (C.17) follows. ☐

Proof of Definition 4.2. The argument here follows closely the proof of (Malekian et al., 2025, Lemma 3.1). Consider two scenarios of our spec6 e noise model:

$$\begin{aligned} \boldsymbol{y}_l =& A_l X_o \boldsymbol{w}_l + \boldsymbol{z}_l, \text{ for } l = 1, \dots, L. \\ \boldsymbol{y}'_l =& A_l X_o \boldsymbol{w}'_l + \boldsymbol{z}'_l, \text{ for } l = 1, \dots, L. \end{aligned} \tag{C.21}$$

where we have independent $\boldsymbol{w}_l \sim \mathcal{N}(0, I_n)$, $\boldsymbol{w}'_l \sim \mathcal{N}(0, I_n)$, $\boldsymbol{z}_l \sim \mathcal{N}(0, \sigma_z I_m)$, and $\boldsymbol{z}'_l \sim \mathcal{N}(0, \sigma'_z I_m)$ with $\sigma'_z > \sigma_z > 0$. We would like to show

$$R_2(\mathcal{C}_k, m, n, \sigma_z) \leq R_2(\mathcal{C}_k, m, n, \sigma'_z). \tag{C.22}$$

Let $\widehat{\boldsymbol{x}}(\boldsymbol{y}'_1, ..., \boldsymbol{y}'_L)$ be any estimator for $\widehat{\boldsymbol{x}}_o$ with observations $\boldsymbol{y}'_1, ..., \boldsymbol{y}'_L$. Note that conditioning on $A_1, ..., A_L$, we have

$$(\boldsymbol{y}'_1, ..., \boldsymbol{y}'_L) \stackrel{d}{=} (\boldsymbol{y}_1 + \boldsymbol{u}_1, ..., \boldsymbol{y}_L + \boldsymbol{u}_L), \tag{C.23}$$

where $\boldsymbol{u}_1, ..., \boldsymbol{u}_L$ are i.i.d. $\mathcal{N}(0, \sqrt{\sigma_z'^2 - \sigma_z^2} \cdot I_m)$. It follows that

$$\mathbb{E}\left[\frac{\left\|\widehat{\boldsymbol{x}}(\boldsymbol{y}'_1, ..., \boldsymbol{y}'_L) - \boldsymbol{x}_o\right\|^2}{n}\right] = \mathbb{E}\left[\frac{\left\|\widehat{\boldsymbol{x}}(\boldsymbol{y}_1 + \boldsymbol{u}_1, ..., \boldsymbol{y}_L + \boldsymbol{u}_L) - \boldsymbol{x}_o\right\|^2}{n}\right]. \tag{C.24}$$

Furthermore, if we let $\mathbb{E}_Y$ denote the conditional expectation $\mathbb{E}_Y[\cdot] := \mathbb{E}[\cdot \mid A_1, ..., A_L, \boldsymbol{y}_1, ..., \boldsymbol{y}_L]$, then by the tower rule and Jensen's inequality we have

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{\boldsymbol{x}}(\boldsymbol{y}_1 + \boldsymbol{u}_1, \dots, \boldsymbol{y}_L + \boldsymbol{u}_L) - \boldsymbol{x}_o\|_2^2\right] \\
=&\mathbb{E}\left[\mathbb{E}_Y\|\widehat{\boldsymbol{x}}(\boldsymbol{y}_1 + \boldsymbol{u}_1, \dots, \boldsymbol{y}_L + \boldsymbol{u}_L) - \boldsymbol{x}_o\|_2^2\right] \\
\geq&\mathbb{E}\left[\|\mathbb{E}_Y\left[\widehat{\boldsymbol{x}}(\boldsymbol{y}_1 + \boldsymbol{u}_1, \dots, \boldsymbol{y}_L + \boldsymbol{u}_L)\right] - \boldsymbol{x}_o\|_2^2\right].
\end{aligned} \tag{C.25}$$

It follows that

$$\begin{aligned}
&\sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\widehat{\boldsymbol{x}}(\boldsymbol{y}_1 + \boldsymbol{u}_1, \dots, \boldsymbol{y}_L + \boldsymbol{u}_L) - \boldsymbol{x}_o\|_2^2\right] \\
\geq &\sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\mathbb{E}_Y\left[\widehat{\boldsymbol{x}}(\boldsymbol{y}_1 + \boldsymbol{u}_1, \dots, \boldsymbol{y}_L + \boldsymbol{u}_L)\right] - \boldsymbol{x}_o\|_2^2\right].
\end{aligned} \tag{C.26}$$

Now, treating $\widehat{\boldsymbol{z}}_o(\boldsymbol{y}_1, \dots, \boldsymbol{y}_L) := \mathbb{E}_Y\left[\widehat{\boldsymbol{x}}(\boldsymbol{y}_1 + \boldsymbol{u}_1, \dots, \boldsymbol{y}_L + \boldsymbol{u}_L)\right]$ as an estimator of $\boldsymbol{x}_o$ using only the information of $\boldsymbol{y}_1, ..., \boldsymbol{y}_L$, we have for every estimator $\widehat{\boldsymbol{z}}(\boldsymbol{y}'_1, ..., \boldsymbol{y}'_L)$ of the second scenario

$$\sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\widehat{\boldsymbol{z}}(\boldsymbol{y}'_1, \dots, \boldsymbol{y}'_L) - \boldsymbol{x}_o\|_2^2\right] \geq \sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\widehat{\boldsymbol{z}}_o(\boldsymbol{y}_1, \dots, \boldsymbol{y}_L) - \boldsymbol{x}_o\|_2^2\right] \geq \inf_{\widehat{\boldsymbol{z}}} \sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\widehat{\boldsymbol{z}}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_L) - \boldsymbol{x}_o\|_2^2\right]. \tag{C.27}$$

Therefore

$$\begin{aligned}
&R_2(\mathcal{C}_k, m, n, \sigma'_z) \\
=&\frac{1}{n} \inf_{\widehat{\boldsymbol{z}}} \sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\widehat{\boldsymbol{z}}(\boldsymbol{y}'_1, \dots, \boldsymbol{y}'_L) - \boldsymbol{x}_o\|_2^2\right] \geq \frac{1}{n} \inf_{\widehat{\boldsymbol{z}}} \sup_{\boldsymbol{x}_o \in \mathcal{C}_k} \mathbb{E}\left[\|\widehat{\boldsymbol{z}}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_L) - \boldsymbol{x}_o\|_2^2\right] \geq R_2(\mathcal{C}_k, m, n, \sigma_z).
\end{aligned} \tag{C.28}$$

□

D Proofs of results of Section 5

D.1 Proof of Definition 5.1

We first establish the lower bound. Let $\boldsymbol{x} \in \mathcal{S}_{\text{sep}}$. Consider all vectors obtained from $\boldsymbol{x}$ by selecting $k/4$ intervals and flip the values of the entries whose indices belong to those intervals; if the value is $\bar{x}+\delta_r$ switch that to $\bar{x}$, and if the value is $\bar{x}$ switch that to $\bar{x}+\delta_r$. Denote the collection of all such vectors by $B(\boldsymbol{x})$. Note that if $\cup_{\boldsymbol{x}\in\mathcal{S}_{\text{sep}}} B(\boldsymbol{x})$ does not cover $\mathcal{X}^{\text{finite}}$, it means that there exists another vector $\boldsymbol{x}\in\mathcal{X}^{\text{finite}}$ that is different from all elements of $\mathcal{S}_{\text{sep}}$ in at least $k/4$ intervals, which is in contradiction with the fact that $\mathcal{S}_{\text{sep}}$ includes **all** such vectors. Hence, we can conclude that

$$\mathcal{X}^{\text{finite}} \subset \cup_{\boldsymbol{x}\in\mathcal{S}_{\text{sep}}} B(\boldsymbol{x}). \tag{D.1}$$

Note that

$$B(\boldsymbol{x}) = \binom{N_{\text{div}}}{k/4} \tag{D.2}$$

Combining (5.12), (D.1), and (D.2) and assuming that r is the size of $\mathcal{S}_{\text{sep}}$ we have

$$r \geq \frac{\binom{N_{\text{div}}}{k}}{\binom{N_{\text{div}}}{k/4}}. \tag{D.3}$$

Using the following classical bounds for $\binom{n}{k}$

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k, \tag{D.4}$$

we have

$$r \geq \frac{\binom{N_{\text{div}}}{k/2}}{\binom{N_{\text{div}}}{k/4}} \geq \frac{(\frac{N_{\text{div}}}{k/2})^{\frac{k}{2}}}{(\frac{eN_{\text{div}}}{k/4})^{\frac{k}{4}}} = (\frac{N_{\text{div}}}{ek})^{\frac{k}{4}}. \tag{D.5}$$

We get the desired upper bound by combining (5.12) and (D.4) with $r \leq |\mathcal{X}^{\text{finite}}|$, as $\mathcal{S}_{\text{sep}} \subset \mathcal{X}^{\text{finite}}$.

D.2 Proof of Definition 5.2

As we discussed in Section 5.3 we have

$$\mathbb{P}_{\boldsymbol{x}} \sim \otimes_{l=1}^{L} N(\mathbf{0}, \Sigma_l^{-1}(\boldsymbol{x})) = N(\mathbf{0}, \boldsymbol{\Sigma}^{-1}(\boldsymbol{x})) \tag{D.6}$$

where

$$\Sigma_l = \Sigma_l(\boldsymbol{x}) := (\sigma_z^2 I_m + A_l X^2 A_l^\top)^{-1} \tag{D.7}$$

and

$$\boldsymbol{\Sigma}(\boldsymbol{x}) := \operatorname{diag}\left(\Sigma_1(\boldsymbol{x}), ..., \Sigma_L(\boldsymbol{x})\right). \tag{D.8}$$

Using Definition 4.4, we condition on $A_1, ..., A_L$, with $\Lambda_1 = \boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}, \Lambda_2 = \boldsymbol{\Sigma}(\boldsymbol{x}_j)^{-1}$, to get

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}_i} \parallel \mathbb{P}_{\boldsymbol{x}_j}) =& \frac{1}{2}\left[\log\frac{\det \boldsymbol{\Sigma}(\boldsymbol{x}_j)^{-1}}{\det \boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}} - mL + \operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}\right)\right] \\
=& \frac{1}{2}\left[\log\det \boldsymbol{\Sigma}(\boldsymbol{x}_j)^{-1}\boldsymbol{\Sigma}(\boldsymbol{x}_i) + \operatorname{Tr}\left(\left[\boldsymbol{\Sigma}(\boldsymbol{x}_j) - \boldsymbol{\Sigma}(\boldsymbol{x}_i)\right]\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}\right)\right] \\
=& \frac{1}{2}\left[-\log\det \boldsymbol{\Sigma}(\boldsymbol{x}_j)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1} + \operatorname{Tr}\left(\left[\boldsymbol{\Sigma}(\boldsymbol{x}_j) - \boldsymbol{\Sigma}(\boldsymbol{x}_i)\right]\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}\right)\right].
\end{aligned} \tag{D.9}$$

In order to find an upper bound for the KL divergence, we use the mean value theorem. By applying the mean value theorem to $\log\det \boldsymbol{\Sigma}(\boldsymbol{x}_j)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}$ and defining λ_q as the q^{th} eigenvalue of the matrix $\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}$, we obtain

$$\begin{aligned}
-\log\det \boldsymbol{\Sigma}(\boldsymbol{x}_j)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1} &= -\log\det\left[\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{x}_j)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1/2}\right]\\
&= -\log\det\left[\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}+I_{mL}\right]\\
&= -\sum_{i=1}^{mL}\log(1+\lambda_q) \overset{(a)}{=} -\sum_{i=1}^{mL}\left(\lambda_q-\frac{\lambda_q^2}{2(1+\lambda_q')^2}\right)\\
&= -\operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\right)+\sum_{i=1}^{mL}\frac{\lambda_q^2}{2(1+\lambda_q')^2},
\end{aligned} \tag{D.10}$$

where to obtain (a) we have used the Taylor expansion for $\log(1+\lambda_q)$, and defined λ_q' as a point between zero and λ_q. Note that this eigenvalue can be negative. Since we have $\operatorname{Tr}\left(\left[\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right]\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}\right) = \operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\right)$, by combining (D.9) and (D.10) we obtain

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}_i}\parallel\mathbb{P}_{\boldsymbol{x}_j}) = \sum_{i=1}^{mL}\frac{\lambda_q^2}{4(1+\lambda_q')^2}. \tag{D.11}$$

Since λ_q' can be a negative number, in order to obtain a useful upper bound for $\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}_i}\parallel\mathbb{P}_{\boldsymbol{x}_j})$ as is required by Fano's inequality we need to find an upper bound for $|\lambda_q'|$. But since this quantity is between zero and λ_q we can bound $|\lambda_q|$ instead. We have

$$\begin{aligned}
&\left|\lambda_q\left(\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\right)\right|\\
\le&\frac{\sigma_{\max}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right)}{\sigma_{\min}(\boldsymbol{\Sigma}(\boldsymbol{x}_i))}\\
\overset{(a)}{\le}&\frac{\sigma_{\max}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)^{-1}-\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}\right)}{\sigma_{\min}(\boldsymbol{\Sigma}(\boldsymbol{x}_i))\sigma_{\min}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}\right)\sigma_{\min}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)^{-1}\right)}\\
=&\frac{\max_{1\le l\le L}\left[\sigma_{\max}(\sigma_z^2 I_m+A_lX_i^2A_l^\top)\right]\max_{1\le l\le L}\left[\sigma_{\max}(A_lX_i^2A_l^\top-A_lX_j^2A_l^\top)\right]}{\sigma_{\min}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}\right)\sigma_{\min}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)^{-1}\right)}.\\
\le&\frac{\left[\sigma_z^2+x_{\max}^2\max_{1\le l\le L}\lambda_{\max}(A_lA_l^\top)\right]\max_{1\le l\le L}\lambda_{\max}(A_lA_l^\top)\|\boldsymbol{x}_i{}^2-\boldsymbol{x}_j^2\|_\infty}{\left(\sigma_z^2+x_{\min}^2\min_{1\le l\le L}\lambda_{\min}(A_lA_l^\top)\right)^2},
\end{aligned} \tag{D.12}$$

where (a) follows from Definition 4.11.

As we discussed before we would like to use our upper bounds for $\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}_i}\parallel\mathbb{P}_{\boldsymbol{x}_j})$ for the Fano's inequality. Hence, in the rest of the proof, we assume that $\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{S}_{\text{sep}}$. This implies that

$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty \le \delta_r$. Hence, we can simplify (D.12) in the following way

$$\begin{aligned}\left|\lambda_q\left(\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right)\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-\frac{1}{2}}\right)\right| \le& \frac{\left[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}\right]\cdot E_{\max}}{\left(\sigma_z^2 + x_{\min}^2 \cdot E_{\min}\right)^2}\cdot 2x_{\max}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty \quad \text{(D.13)}\\ \le& \frac{\left[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}\right]\cdot E_{\max}}{\left(\sigma_z^2 + x_{\min}^2 \cdot E_{\min}\right)^2}\cdot 2x_{\max}\delta_r. \quad \text{(D.14)}\end{aligned}$$

Hence, choosing δ_r such that $\frac{\left[\sigma_z^2+x_{\max}^2\cdot E_{\max}\right]\cdot E_{\max}}{\left(\sigma_z^2+x_{\min}^2\cdot E_{\min}\right)^2}\cdot 2x_{\max}\delta_r < \frac{1}{2}$, we use (D.11) to get

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_i \parallel \mathbb{P}_j) \le& \sum_{i=1}^{mL}\frac{\lambda_q^2}{4(1+\lambda_q')^2} \le \sum_{i=1}^{mL}\lambda_q^2\\ =&\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}\left[\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right]\boldsymbol{\Sigma}(\boldsymbol{x}_i)^{-1}\left[\boldsymbol{\Sigma}(\boldsymbol{x}_j)-\boldsymbol{\Sigma}(\boldsymbol{x}_i)\right]\right) \\ \le& \frac{\left(\sigma_z^2 + x_{\max}^2 \max_{1\le l\le L}\lambda_{\max}(A_lA_l^\top)\right)^2}{\left(\sigma_z^2 + x_{\min}^2 \min_{1\le l\le L}\lambda_{\min}(A_lA_l^\top)\right)^4}\sum_{l=1}^{L}\left\|A_l(X_i^2 - X_j^2)A_l^\top\right\|_{\mathrm{HS}}^2,\end{aligned} \tag{D.15}$$

where to obtain the last inequality we have used Lemma 4.14.

E Proof of technical results

E.1 Proof of Lemma 7.1

As stated from the beginning of Subsection 7.1, without loss of generality, we may assume $\mathcal{C} \in \mathcal{F}_{a,b,k,n}$ with $a = x_{\max} - x_{\min}$ and $b = 1$. Namely

$$|N_{\delta_{\mathrm{net}}}(\mathcal{C}_k)| \le \left(\frac{(x_{\max}-x_{\min})n}{\delta_{\mathrm{net}}}\right)^k. \tag{E.1}$$

Fix a general $\widetilde{\boldsymbol{x}}_o \in N_{\delta_{\mathrm{net}}}(\mathcal{C}_k)$. For $1 \le l \le L$, define the matrices

$$\begin{aligned}A_{\sigma_z,l} = [\sigma_z I_m \ A_l X_o] \in \mathbb{R}^{m\times(m+n)}, \quad B_l = A_{\sigma_z,l}^\top\left(\Sigma_l(\widetilde{\boldsymbol{x}}_o) - \Sigma_l(\boldsymbol{x}_o)\right)A_{\sigma_z,l} \in \mathbb{R}^{(m+n)\times(m+n)},\\ \boldsymbol{A}_{\sigma_z} = \mathrm{diag}(A_{\sigma_z,1},\dots,A_{\sigma_z,L}) \in \mathbb{R}^{mL\times(m+n)L}, \quad \boldsymbol{B} = \mathrm{diag}(B_1,\dots,B_L) \in \mathbb{R}^{L(m+n)\times L(m+n)}.\end{aligned} \tag{E.2}$$

In view of the notations (7.6) and (7.14), the above display provides us with the following identities

$$\boldsymbol{A}_{\sigma_z}\boldsymbol{A}_{\sigma_z}^\top = \boldsymbol{\Sigma}_o^{-1}, \quad \boldsymbol{B} = \boldsymbol{A}_{\sigma_z}^\top\left(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\boldsymbol{A}_{\sigma_z}. \tag{E.3}$$

Define the $L(m+n)$ dimensional vector, $\vec{\boldsymbol{w}}^\top = [\boldsymbol{z}_1^\top/\sigma_z, \boldsymbol{w}_1^\top, \dots, \boldsymbol{z}_L^\top/\sigma_z, \boldsymbol{w}_L^\top]$. It follows that

$$\vec{\boldsymbol{y}}^\top\left(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\vec{\boldsymbol{y}} = \vec{\boldsymbol{w}}^\top\boldsymbol{B}\vec{\boldsymbol{w}}. \tag{E.4}$$

Then, conditioning on $A_1, ..., A_L$, by the Hanson-Wright inequality (Definition 4.8), we have

$$\mathbb{P}\left[|\vec{\boldsymbol{w}}^\top\boldsymbol{B}\vec{\boldsymbol{w}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o))| > t\right] \le 2\exp\left(-c\min\left(\frac{t^2}{4\|\boldsymbol{B}\|_{\mathrm{HS}}^2}, \frac{t}{2\|\boldsymbol{B}\|_2}\right)\right). \tag{E.5}$$

We simplify the above using upper bounds on $\|\boldsymbol{B}\|_{\text{HS}}, \|\boldsymbol{B}\|_2$. In view of (7.12) we bound $\|\boldsymbol{B}\|_2$ as

$$\begin{aligned}
&\|\boldsymbol{B}\|_2 = \max\left(\|B_l\|_2\right)_{l=1}^L \\
&\le \left(\sigma_z^2 + x_{\max}^2 \max_{1\le l\le L} \sigma_{\max}(A_l)^2\right) \max_{1\le l\le L} \sigma_{\max}\left(\Sigma_l(\widetilde{\boldsymbol{x}}_o) - \Sigma_l(\boldsymbol{x}_o)\right) \\
&\le \left(\sigma_z^2 + x_{\max}^2 E_{\max}\right) \max_{1\le l\le L} \left(\sigma_{\max}(\Sigma_l(\widetilde{\boldsymbol{x}}_o)) + \sigma_{\max}(\Sigma_l(\boldsymbol{x}_o))\right) \\
&= \left(\sigma_z^2 + x_{\max}^2 E_{\max}\right) \max_{1\le l\le L} \left(\{\sigma_{\min}(\Sigma_l(\widetilde{\boldsymbol{x}}_o)^{-1})\}^{-1} + \{\sigma_{\min}(\Sigma_l(\boldsymbol{x}_o)^{-1})\}^{-1}\right) \le C_{n,m,\sigma_z},
\end{aligned} \tag{E.6}$$

where the last inequality followed by noting that on the event $\mathcal{S}$ in (7.3) we have for each $\boldsymbol{x} = \boldsymbol{x}_o, \widetilde{\boldsymbol{x}}_o$

$$\sigma_{\min}(\Sigma_l(\boldsymbol{x})^{-1}) = \sigma_z^2 + \sigma_{\min}(A_l X^2 A_l^\top) \ge \sigma_z^2 + x_{\min}^2 \sigma_{\min}(A_l A_l^\top) \ge \sigma_z^2 + x_{\min}^2 E_{\min}, \quad 1 \le l \le L.$$

Next, using the identity $\|\boldsymbol{B}\|_{\text{HS}}^2 = \text{Tr}(\boldsymbol{B}^2)$ we get

$$\|\boldsymbol{B}\|_{\text{HS}}^2 = \text{Tr}\left[\boldsymbol{\Sigma}_o^{-1}\left(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\boldsymbol{\Sigma}_o^{-1}\left(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\right]. \tag{E.7}$$

which we have defined as $\mathscr{X}$. Hence, conditioned on the event $\mathcal{S}$ in (7.3), we simplify (E.5) to get

$$\mathbb{P}\left[|\vec{\boldsymbol{w}}^\top \boldsymbol{B}\vec{\boldsymbol{w}} - \text{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o))| > t \,\Big|\, \mathcal{S}\right] \le 2\exp\left(-c\min\left(\frac{t^2}{4\mathscr{X}}, \frac{t}{2C_{n,m,\sigma_z}}\right)\right). \tag{E.8}$$

To extend the above probability statement for all possible choice of $\widetilde{\boldsymbol{x}}_o \in N_{\delta_{\text{net}}}(\mathcal{C}_k)$ we use an union bound argument.

The above implies that by choosing $t = b_1\sqrt{\mathscr{X}} + b_1'$ as defined in the result statement, we get

$$\begin{aligned}
&\mathbb{P}\left[|\vec{\boldsymbol{w}}^\top \boldsymbol{B}\vec{\boldsymbol{w}} - \text{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o))| > t \text{ for all } \widetilde{\boldsymbol{x}}_o \in N_{\delta_{\text{net}}}(\mathcal{C}_k) \,\Big|\, \mathcal{S}\right] \\
&\le e^{k\log(3x_{\max}\sqrt{n}/\delta_{\text{net}})} \exp\left(-c\min\left\{\frac{b_1^2}{4}, \frac{b_1'}{2C_{n,m,\sigma_z}}\right\}\right) \le e^{-\tilde{c}k\log(3x_{\max}\sqrt{n}/\delta_{\text{net}})},
\end{aligned} \tag{E.9}$$

for some constant $\tilde{c} > 0$. In view of (7.3) and $\mathbb{P}[A] \le \mathbb{P}[A|\mathcal{S}] + \mathbb{P}[\mathcal{S}^c]$, the above display implies

$$\mathbb{P}\left[|\vec{\boldsymbol{w}}^\top \boldsymbol{B}\vec{\boldsymbol{w}} - \text{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o))| > t \text{ for all } \widetilde{\boldsymbol{x}}_o \in N_{\delta_{\text{net}}}(\mathcal{C}_k)\right] \le e^{-\tilde{c}k\log\left(\frac{3x_{\max}\sqrt{n}}{\delta_{\text{net}}}\right)} + Le^{-cn}.$$

E.2 Proof of Lemma 7.2

Our entire argument is conditioning on the high-probability event $\mathcal{S}$ in (7.3). We will explain at the end of the section how the conditioning is removed to get the final result. In view of Lemma 4.11 we first note that

$$\begin{aligned}
\sigma_{\max}(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o) &\le \frac{\sigma_{\max}\left(\widehat{\boldsymbol{\Sigma}}_o^{-1} - \widetilde{\boldsymbol{\Sigma}}_o^{-1}\right)}{\sigma_{\min}\left(\widehat{\boldsymbol{\Sigma}}_o^{-1}\right)\sigma_{\min}\left(\widetilde{\boldsymbol{\Sigma}}_o^{-1}\right)} \le \frac{\max_{1\le l\le L}\left[\sigma_{\max}(A_l(\widehat{X}_o^2 - \widetilde{X}_o^2)A_l^\top)\right]}{\sigma_{\min}\left(\widehat{\boldsymbol{\Sigma}}_o^{-1}\right)\sigma_{\min}\left(\widetilde{\boldsymbol{\Sigma}}_o^{-1}\right)} \\
&\le \frac{\max_{1\le l\le L}\lambda_{\max}(A_l A_l^\top)\|\widehat{\boldsymbol{x}}_o^2 - \widetilde{\boldsymbol{x}}_o^2\|_\infty}{\left(\sigma_z^2 + x_{\min}^2 \min_{1\le l\le L}\lambda_{\min}(A_l A_l^\top)\right)^2} \le \frac{2E_{\max}x_{\max}\delta_{\text{net}}}{(\sigma_z^2 + x_{\min}^2 \cdot E_{\min})^2}.
\end{aligned} \tag{E.10}$$

In view of $\sigma_{\max}(\mathbf{\Sigma}_o^{-1}) \le \sigma_z^2 + x_{\max}^2 E_{\max}$, for any $1 \le q \le mL$ we use the last display to get

$$\left|\lambda_q(\mathbf{\Sigma}_o^{-1}(\widehat{\mathbf{\Sigma}}_o - \widetilde{\mathbf{\Sigma}}_o))\right| \le \frac{\sigma_{\max}\left(\widetilde{\mathbf{\Sigma}}_o - \widehat{\mathbf{\Sigma}}_o\right)}{\sigma_{\min}(\mathbf{\Sigma}_o)} \le \frac{[\sigma_z^2 + x_{\max}^2 E_{\max}] \cdot E_{\max}}{\left(\sigma_z^2 + x_{\min}^2 \cdot E_{\min}\right)^2} \cdot 2x_{\max}\delta_{\text{net}}, \tag{E.11}$$

with probability $1 - Le^{-cn}$. Summing up over $1 \le q \le mL$ and using triangular inequality, we have

$$\left|\mathrm{Tr}[\mathbf{\Sigma}_o^{-1}(\widehat{\mathbf{\Sigma}}_o - \widetilde{\mathbf{\Sigma}}_o)]\right| \le mL \cdot \frac{[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}] \cdot E_{\max}}{\left(\sigma_z^2 + x_{\min}^2 \cdot E_{\min}\right)^2} \cdot 2x_{\max}\delta_{\text{net}}. \tag{E.12}$$

To bound the term $|\vec{\boldsymbol{y}}^\top(\widehat{\mathbf{\Sigma}}_o - \widetilde{\mathbf{\Sigma}}_o)\vec{\boldsymbol{y}}|$, we again use (E.10) to get

$$\left|\vec{\boldsymbol{y}}^\top \left(\widehat{\mathbf{\Sigma}}_o - \widetilde{\mathbf{\Sigma}}_o\right) \vec{\boldsymbol{y}}\right| \le \sigma_{\max}\left(\widehat{\mathbf{\Sigma}}_o - \widetilde{\mathbf{\Sigma}}_o\right) \vec{\boldsymbol{y}}^\top \vec{\boldsymbol{y}} \le \frac{2E_{\max}x_{\max}\delta_{\text{net}}}{\left(\sigma_z^2 + x_{\min}^2 \cdot E_{\min}\right)^2} \vec{\boldsymbol{y}}^\top \vec{\boldsymbol{y}}. \tag{E.13}$$

We use the following lemma to bound $\vec{\boldsymbol{y}}^\top \vec{\boldsymbol{y}}$. A proof is provided at the end of this section.

Lemma E.1. For $t > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\vec{\boldsymbol{y}}^\top \vec{\boldsymbol{y}} \ge mL\left[E_{\max}x_{\max}^2 + \sigma_z^2\right] + t\right) \\
&\le 2\exp\left(-c\min\left(\frac{t^2}{4mL\left(\sigma_z^2 + x_{\max}^2 \cdot E_{\max}\right)^2}, \frac{t}{2\left(\sigma_z^2 + x_{\max}^2 \cdot E_{\max}\right)}\right)\right) + Le^{-cn}.
\end{aligned} \tag{E.14}$$

We choose $t = CmL\left[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}\right]$ for a large constant $C > 0$. Then, in view of Lemma E.1, we we continue (E.13) to get that with probability $1 - Le^{-cn} - \exp(-cmL)$ the following holds

$$\left|\vec{\boldsymbol{y}}^\top \left(\widehat{\mathbf{\Sigma}}_o - \widetilde{\mathbf{\Sigma}}_o\right) \vec{\boldsymbol{y}}\right| \le \frac{2x_{\max}\delta_{\text{net}}E_{\max}}{\left(\sigma_z^2 + x_{\min}^2 \cdot E_{\min}\right)^2} \cdot 2mL\left[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}\right]. \tag{E.15}$$

Together with (E.12), we conclude that with probability $1 - Le^{-n} - e^{-cmL}$ that

$$\left|\vec{\boldsymbol{y}}^\top(\mathbf{\Sigma}_o - \widehat{\mathbf{\Sigma}}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\mathbf{\Sigma}_o^{-1}(\mathbf{\Sigma}_o - \widehat{\mathbf{\Sigma}}_o))\right| \le 2mLx_{\max}\delta_{\text{net}} \cdot \frac{[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}] \cdot E_{\max}}{\left(\sigma_z^2 + x_{\min}^2 \cdot E_{\min}\right)^2}. \tag{E.16}$$

As $mLx_{\max}\delta_{\text{net}} \cdot \frac{[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}] \cdot E_{\max}}{\left(\sigma_z^2 + x_{\min}^2 \cdot E_{\min}\right)^2} \le mL\delta_{\text{net}}(C_{n,m,\sigma_z})^2$ we get the desired result.

Proof of Definition E.1. Define

$$\begin{aligned}
M_l = A_{\sigma_z,l}^\top A_{\sigma_z,l}, \quad l \in [L], \quad \boldsymbol{M} = \mathrm{diag}(M_1, \ldots, M_L) \in \mathbb{R}^{L(m+n)\times L(m+n)} \\
\vec{\boldsymbol{w}}^\top = [\boldsymbol{z}_1^\top/\sigma_z, \boldsymbol{w}_1^\top, \ldots, \boldsymbol{z}_L^\top/\sigma_z, \boldsymbol{w}_L^\top]
\end{aligned} \tag{E.17}$$

Then we can write $\vec{\boldsymbol{y}}^\top \vec{\boldsymbol{y}} = \vec{\boldsymbol{w}}^\top \boldsymbol{M} \vec{\boldsymbol{w}}^\top$. To obtain a tail bound, we first recall $\vec{\boldsymbol{y}} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L]$, and observe that by conditioning on the event $\mathcal{S}$ as in (7.3), we have for each $l = 1, \ldots, L$,

$$\mathbb{E}[\boldsymbol{y}_l^\top \boldsymbol{y}_l \mid \mathcal{S}] = \mathrm{Tr}\left[X_o A_l^\top A_l X_o\right] + m\sigma_z^2 \le m\left(\lambda_{\max}(A_l^\top A_l)x_{\max}^2 + \sigma_z^2\right) \le m\left(E_{\max}x_{\max}^2 + \sigma_z^2\right). \tag{E.18}$$

Then, by the Hanson-Wright inequality (Definition 4.8), we have conditioned on the event $\mathcal{S}$ that,

$$\mathbb{P}\left(\vec{\boldsymbol{y}}^\top \vec{\boldsymbol{y}} > mL\left(E_{\max} x_{\max}^2 + \sigma_z^2\right) + t \middle| \mathcal{S} \right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\|\boldsymbol{M}\|_{\mathrm{HS}}^2}, \frac{t}{\|\boldsymbol{M}\|_2} \right) \right), \tag{E.19}$$

for some constant $c > 0$. To bound the terms $\|\boldsymbol{M}\|_{\mathrm{HS}}, \|\boldsymbol{M}\|_2$, we note that on the event $\mathcal{S}$,

$$\begin{aligned} \|\boldsymbol{M}\|_2 &= \max_{1\leq l\leq L} \|M_l\|_2 \leq \sigma_z^2 + x_{\max}^2 \max_{1\leq l\leq m} \lambda_{\max}(A_l^\top A_l) \leq \sigma_z^2 + x_{\max}^2 E_{\max}, \\ \|\boldsymbol{M}\|_{\mathrm{HS}}^2 &= \sum_{l=1}^{L} \mathrm{Tr}(M_l^2) \leq \sum_{l=1}^{L}\sum_{i=1}^{m} \lambda_i^2(M_l) \leq \sum_{l=1}^{L} m\lambda_{\max}^2(M_l) \leq Lm\left(\sigma_z^2 + x_{\max}^2 E_{\max}\right)^2, \end{aligned} \tag{E.20}$$

where the inequality $\sum_{l=1}^{L} \mathrm{Tr}(M_l^2) \leq \sum_{l=1}^{L}\sum_{i=1}^{m} \lambda_i^2(M_l)$ follows from the fact that $\mathrm{rank}(M_l) = m$ and there are at most m nonzero eigenvalues. Finally we remove the condition on $\mathcal{S}$ by using $\mathbb{P}[A] \leq \mathbb{P}[A|\mathcal{S}] + \mathbb{P}[\mathcal{S}^c]$ with $\mathbb{P}[\mathcal{S}^c] \leq Le^{-cn}$ from (7.4) to get the desired result.

□

E.3 Proof of Lemma 7.3

The proof here is the similar to that of (Zhou et al., 2022, Lemma VIII.4) with minor differences, which we point out below. Recall that,

$$\overline{f}(\widetilde{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o) = \left[-\log\det\widetilde{\boldsymbol{\Sigma}}_o + \mathrm{Tr}\left(\widetilde{\boldsymbol{\Sigma}}_o \boldsymbol{\Sigma}_o^{-1}\right)\right] - \left[-\log\det\boldsymbol{\Sigma}_o + \mathrm{Tr}\left(\boldsymbol{\Sigma}_o\boldsymbol{\Sigma}_o^{-1}\right)\right] \tag{E.21}$$

$$= -\left[\log\det\widetilde{\boldsymbol{\Sigma}}_o - \log\det\boldsymbol{\Sigma}_o\right] + \mathrm{Tr}([\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o]\boldsymbol{\Sigma}_o^{-1}) \tag{E.22}$$

For $1 \leq q \leq mL$, let λ_q denote the q-th eigenvalue of $\boldsymbol{\Sigma}_o^{-\frac{1}{2}}\left(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\boldsymbol{\Sigma}_o^{-\frac{1}{2}}$, and $\lambda_{\max} = \max_{1\leq q\leq mL} |\lambda_q|$. Following the proof strategy of (Zhou et al., 2022, Lemma VIII.4) we can show that

$$\log\frac{\det\widetilde{\boldsymbol{\Sigma}}_o}{\det\boldsymbol{\Sigma}_o} \leq \mathrm{Tr}\left(\boldsymbol{\Sigma}_o^{-\frac{1}{2}}\left(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\boldsymbol{\Sigma}_o^{-\frac{1}{2}}\right) - \frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\right)}{2(1+\widetilde{\lambda}_{\max})^2}. \tag{E.23}$$

We will use the following inequalities to bound $\widetilde{\lambda}_{\max}$, on the event $\mathcal{S}$.

- $\|\boldsymbol{\Sigma}_o^{-1}\|_2 = \max_{1\leq l\leq L} \|A_l X_o^2 A_l^\top\|_2 \leq \sigma_z^2 + x_{\max}^2 \max_{1\leq l\leq L} \sigma_{\max}(A_l A_l^\top) \leq \sigma_z^2 + x_{\max}^2 n$.
- Using $\|E\|^{-1} = \sigma_{\min}(E^{-1})$ for any invertible matrix E, for a constant C, we get

$$\begin{aligned} \|\boldsymbol{\Sigma}_o\|^{-1} &= \sigma_z^2 + \min_{1\leq l\leq L} \sigma_{\min}(A_l X_o^2 A_l^\top) \geq \sigma_z^2 + x_{\min}^2 \min_{1\leq l\leq L} \sigma_{\min}(A_l A_l^\top) \geq \sigma_z^2 + x_{\min}^2 E_{\min} \\ \|\widetilde{\boldsymbol{\Sigma}}_o\|^{-1} &= \sigma_z^2 + \min_{1\leq l\leq L} \sigma_{\min}(A_l \tilde{X}_o^2 A_l^\top) \geq \sigma_z^2 + x_{\min}^2 \min_{1\leq l\leq L} \sigma_{\min}(A_l A_l^\top) \geq \sigma_z^2 + x_{\min}^2 E_{\min}. \end{aligned}$$

Using $\widetilde{\lambda}_{\max} = \|\boldsymbol{\Sigma}_o^{-\frac{1}{2}}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\boldsymbol{\Sigma}_o^{-\frac{1}{2}}\|_2 \leq \|\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\|_2 \|\boldsymbol{\Sigma}_o^{-\frac{1}{2}}\|_2^2 = \|\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\|_2 \|\boldsymbol{\Sigma}_o^{-1}\|_2$, we get

$$\widetilde{\lambda}_{\max} \leq (\|\widetilde{\boldsymbol{\Sigma}}_o\|_2 + \|\boldsymbol{\Sigma}_o\|_2)\|\boldsymbol{\Sigma}_o^{-1}\|_2 \leq 2\frac{\sigma_z^2 + x_{\max}^2 E_{\max}}{\sigma_z^2 + x_{\min}^2 E_{\min}} \leq C_{n,m,\sigma_z}. \tag{E.24}$$

E.4 Proof of Lemma 7.4

For the first term $\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\widetilde{\boldsymbol{x}}_o)$, using an argument similar to the proof of Lemma 7.3 we can show

$$\left|\overline{f}(\widetilde{\boldsymbol{x}}_o) - \overline{f}(\widehat{\boldsymbol{x}}_o)\right| \le \frac{1}{2(1-|\lambda_{\max}|)^2}\mathrm{Tr}\left(\widehat{\boldsymbol{\Sigma}}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o)\widehat{\boldsymbol{\Sigma}}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o)\right) \tag{E.25}$$

where $|\lambda_{\max}|$ is the largest absolute value of eigenvalues of $\widehat{\boldsymbol{\Sigma}}_o^{-\frac{1}{2}}(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o)\widehat{\boldsymbol{\Sigma}}_o^{-\frac{1}{2}}$ and with probability at least $1-Le^{-cn}$. Choosing $\delta_{\text{net}} = \frac{x_{\max}}{n^5}$, and noting that $|\lambda_{\max}| \le \|\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o\|_2\|\widehat{\boldsymbol{\Sigma}}_o^{-1}\|_2$, we get

$$\begin{aligned}
|\lambda_{\max}| &\overset{(a)}{\le} \frac{\sigma_{\max}(\widehat{\boldsymbol{\Sigma}}_o^{-1} - \widetilde{\boldsymbol{\Sigma}}_o^{-1})\|\widehat{\boldsymbol{\Sigma}}_o^{-1}\|_2}{\sigma_{\min}(\widehat{\boldsymbol{\Sigma}}_o^{-1})\sigma_{\min}(\widetilde{\boldsymbol{\Sigma}}_o^{-1})} \\
&\le \frac{\sigma_z^2 + x_{\max}^2 \cdot E_{\max}}{\sigma_z^2 + x_{\min}^2 \cdot E_{\min}}^2 \max_{1\le l\le L}\sigma_{\max}\left(A_l(\widetilde{X}_o^2 - \widehat{X}_o^2)A_l^\top\right) \le C_{n,m,\sigma_z}\cdot x_{\max} n\delta_{\text{net}}, \ll 1
\end{aligned}$$

where in (a) we have used Definition 4.11.

E.5 Proof of Lemma 7.5

Pick an $\widetilde{\boldsymbol{x}}_o \in N_{\delta_{\text{net}}}(\mathcal{C}_k)$. Define $\boldsymbol{d} = \widetilde{\boldsymbol{x}}_o^2 - \boldsymbol{x}_o^2$ and $D = \mathrm{diag}(\boldsymbol{d})$. In view of the above, we have

$$\|\boldsymbol{d}\|_\infty \le x_{\max}^2, \quad \|\boldsymbol{d}^2\|_2 \le 4x_{\max}^2\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2. \tag{E.26}$$

Then using (4.12), as $n \ge 4m$, conditional on the event $\mathcal{S}$ in (7.3), we get

$$\begin{aligned}
&\mathbb{P}\left(\left[\sum_{l=1}^L \|A_l(\widetilde{X}_o^2 - X_o^2)A_l^\top\|_{\text{HS}}^2 < Lm(m-1)\sum_{i=1}^n d_i^2 - t\right]\middle|\mathcal{S}\right) \\
&\le \exp\left(-c\cdot\min\left(\frac{t^2}{K^4x_{\max}^4\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2 mLn^2}, \frac{t}{K^2x_{\max}^2 n}\right)\right) \\
&+ 2m\exp\left(-c\min\left(\frac{t^2}{K^4Lm^3x_{\max}^4\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2}, \frac{t}{K^2mx_{\max}^2}\right)\right).
\end{aligned} \tag{E.27}$$

Choose $t_o = C\log m\left(x_{\max}^2\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2 n\sqrt{mL}\sqrt{k\log\frac{(x_{\max}-x_{\min})n}{\delta_{\text{net}}}} + x_{\max}^4 nk\log\frac{(x_{\max}-x_{\min})n}{\delta_{\text{net}}}\right)$ for a large constant C to be chosen later. Then, the above display implies for a large constant C_1

$$\mathbb{P}\left[\left[\sum_{l=1}^L \|A_l(\widetilde{X}_o^2 - X_o^2)A_l^\top\|_{\text{HS}}^2 < Lm(m-1)\sum_{i=1}^n d_i^2 - t_0\right]\middle|\mathcal{S}\right] \le e^{-C_1\log\frac{(x_{\max}-x_{\min})n}{\delta_{\text{net}}}}$$

Then using an union bound over the total possible choices of $\widetilde{\boldsymbol{x}}_o \in N_{\delta_{\text{net}}}(\mathcal{C}_k)$, with $|N_{\delta_{\text{net}}}(\mathcal{C}_k)| \le \left(\frac{3x_{\max}\sqrt{n}}{\delta_{\text{net}}}\right)$ as in (E.1), we get that as C_1 is large enough,

$$\mathbb{P}\left[\left[\sum_{l=1}^L \|A_l(\widetilde{X}_o^2 - X_o^2)A_l^\top\|_{\text{HS}}^2 < Lm(m-1)\sum_{i=1}^n d_i^2 - t_o\right] \text{ for any } \widetilde{\boldsymbol{x}}_o \in N_{\delta_{\text{net}}}(\mathcal{C}_k)\middle|\mathcal{S}\right] \le e^{-2k\log n}.$$

Then using $\mathbb{P}[B] \le \mathbb{P}[B|\mathcal{S}] + \mathbb{P}[\mathcal{S}^c]$, with B as the event $\sum_{l=1}^L \|A_l(\widetilde{X}_o^2 - X_o^2)A_l^\top\|_{\mathrm{HS}}^2 < Lm(m-1)\sum_{i=1}^n d_i^2 - t_o$, and the fact $\mathbb{P}[\mathcal{S}^c] \le Le^{-cn}$ as in (7.4) we conclude that $\mathbb{P}[B] \le e^{-2k\log n} + Le^{-cn}$. This implies our the desired result.

F Comparison to the fixed forward operator model

F.1 Proof of Definition 2.14

The proofs in this section uses a similar approach to the proofs in Section 7, with the key modifications $A_1 = \cdots = A_L = A$. To proceed with the details, we first define the notations we use throughout the section, and then point out the differences with the proofs in Section 7. The proof of the related technical results, particularly Lemma F.1, Lemma F.2, Lemma F.3, Lemma F.4, Lemma F.5, follow from the proofs of the results in Section 7 by noting that the related proof in the multilook setting uses bounds on the singular values of the sensor matrices $A_1, \dots, A_L$, for which we used a common bound that also holds true for the fixed measurement matrix A. In addition, of results in the fixed sensor case provide guarantees with a higher probability as we do not need to have a uniform control of the singular values of the sensor matrix, as required in the independent multilook setup. This will also hold true for the subsequent results. We omit the technical details.

Similar to before, we will show the desired upper bound is achieved by the maximum likelihood estimator, over is the class of all possible signals $\mathcal{C}$

$$\widehat{\boldsymbol{x}}_o = \arg\min_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = L\log\det\left(\sigma_z^2 I_m + AX^2A^\top\right) + \sum_{l=1}^{L} \boldsymbol{y}_l^\top \left(\sigma_z^2 I_m + AX^2A^\top\right)^{-1} \boldsymbol{y}_l \quad \text{(F.1)}$$

For the entirety of the analysis in this section, we will restrict ourselves to the following event

$$\mathcal{S}_{\text{fix}} = \left\{A : \sigma_{\min}(AA^\top) \ge E_{\min}, \quad \sigma_{\max}(AA^\top) \le E_{\max}\right\}, \quad \text{(F.2)}$$

where $E_{\max}$ and $E_{\min}$ according to the following rules

- Case I ($n \ge 4m$): We will choose $E_{\max} = \frac{9}{4}(\sqrt{n}+\sqrt{m})^2$ and $E_{\min} = \frac{1}{4}(\sqrt{n}-\sqrt{m})^2$. In that case the event $\mathcal{S}_{\text{fix}}$ satisfies $\mathcal{E}_{\text{sing}} \subseteq \mathcal{S}_{\text{fix}}$, where $\mathcal{E}_{\text{sing}}$ is given as in (4.8). This implies $\mathbb{P}\left[\mathcal{S}_{\text{fix}}\right] \ge \mathbb{P}[\mathcal{E}_{\text{sing}}] \ge 1 - e^{-cn}$ for some constant $c > 0$.
- Case II ($n < 4m$): We will choose $E_{\max} = \frac{9}{4}(\sqrt{n}+\sqrt{m})^2$ and $E_{\min} = 0$. In that case the event $\mathcal{S}_{\text{fix}}$ satisfies $\mathcal{E}_{\text{maxsing}} \subseteq \mathcal{S}_{\text{fix}}$, where $\mathcal{E}_{\text{maxsing}}$ is given as in (4.6). This implies $\mathbb{P}\left[\mathcal{S}_{\text{fix}}\right] \ge \mathbb{P}[\mathcal{E}_{\text{maxsing}}] \ge 1 - e^{-cn}$ for some constant $c > 0$.

Consider the following notations for simplifying the presentation. Let Σ be the inverse of the covariance matrix $\mathbb{E}\left[\boldsymbol{y}_l\boldsymbol{y}_l^\top | A\right]$ given by

$$\Sigma = \Sigma(\boldsymbol{x}) := (\sigma_z^2 I_m + AX^2A^\top)^{-1} \quad \text{(F.3)}$$

Define the vector $\vec{\boldsymbol{y}} \in \mathbb{R}^{mL}$ and block-diagonal matrix $\boldsymbol{\Sigma}(\boldsymbol{x}) \in \mathbb{R}^{mL\times mL}$ as the collection of all the observations and the inverse covariance matrices over different looks

$$\vec{\boldsymbol{y}}^\top := (\boldsymbol{y}_1^\top, ..., \boldsymbol{y}_L^\top)^\top, \quad \boldsymbol{\Sigma}(\boldsymbol{x}) := \operatorname{diag}\left(\Sigma(\boldsymbol{x}), ..., \Sigma(\boldsymbol{x})\right), \quad \boldsymbol{\Sigma}_o = \boldsymbol{\Sigma}(\boldsymbol{x}_o), \quad \widehat{\boldsymbol{\Sigma}}_o = \boldsymbol{\Sigma}(\hat{\boldsymbol{x}}_o). \quad \text{(F.4)}$$

In view of the above notations, we can rewrite the negative log-likelihood in (F.1) as

$$f(\boldsymbol{x}) = -\log\det(\boldsymbol{\Sigma}(\boldsymbol{x})) + \vec{\boldsymbol{y}}^\top \boldsymbol{\Sigma}(\boldsymbol{x})\vec{\boldsymbol{y}} \tag{F.5}$$

Now we proceed with the proof. Since $\widehat{\boldsymbol{x}}_o$ is the minimizer from (F.1), we have

$$f(\widehat{\boldsymbol{x}}_o) \leq f(\boldsymbol{x}_o). \tag{F.6}$$

For a fixed $\boldsymbol{x}$, define $\overline{f}(\boldsymbol{x})$ as the function of conditional expectation of $f(\boldsymbol{x})$ given A

$$\overline{f}(\boldsymbol{x}) := \mathbb{E}[f(\boldsymbol{x}) \mid A] = -\log\det\boldsymbol{\Sigma}(\boldsymbol{x}) + \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{x})\boldsymbol{\Sigma}(\boldsymbol{x}_o)^{-1}\right). \tag{F.7}$$

Simplifying the expression for $f(\widehat{\boldsymbol{x}}_o) - f(\boldsymbol{x}_o)$, with the above notations, we get

$$\begin{aligned}
&f(\widehat{\boldsymbol{x}}_o) - f(\boldsymbol{x}_o)\\
&= \vec{\boldsymbol{y}}^\top\left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\vec{\boldsymbol{y}} - \mathrm{Tr}\left[\boldsymbol{\Sigma}_o^{-1}\left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\right] + \mathrm{Tr}\left[\boldsymbol{\Sigma}_o^{-1}\left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\right] - \log\det(\widehat{\boldsymbol{\Sigma}}_o) + \log\det(\boldsymbol{\Sigma}_o)\\
&= \vec{\boldsymbol{y}}^\top\left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\vec{\boldsymbol{y}} - \mathrm{Tr}\left[\boldsymbol{\Sigma}_o^{-1}\left(\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o\right)\right] + \overline{f}(\hat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o).
\end{aligned} \tag{F.8}$$

Therefore, in view of (F.6) we get

$$\vec{\boldsymbol{y}}^\top\left(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o\right)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)) \geq \overline{f}(\hat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o). \tag{F.9}$$

Our following approach is to find an upper bound for the left side in terms of $\|\widehat{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2$ and a lower bound for the right side in terms of $\|\widehat{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2$, and simplify the inequality to get an upper bound for $\|\widehat{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2$. We will use the following notation, similar to (7.12), for a constant $c > 0$

$$C_{n,m,\sigma_z} = C(n, m, \sigma_z, x_{\max}, x_{\min}) = c\frac{\sigma_z^2 + x_{\max}^2 E_{\max}}{\sigma_z^2 + x_{\min}^2 E_{\min}}. \tag{F.10}$$

Establishing an upper bound on $\vec{\boldsymbol{y}}^\top(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o))$: We use the same δ-net argument as in Section 7. Consider a δ-net of the set $\mathcal{C}_k$, denoted by $N_{\delta_{\text{net}}}(\mathcal{C}_k)$, with the choice of δ_{net} to be discussed later. Define $\widetilde{\boldsymbol{x}}_o$ as the closest vector in $N_{\delta_{\text{net}}}(\mathcal{C}_k)$ to $\boldsymbol{x}_o$, i.e.,

$$\widetilde{\boldsymbol{x}}_o = \mathrm{argmin}_{\boldsymbol{x}\in N_{\delta_{\text{net}}}(\mathcal{C}_k)} \|\widehat{\boldsymbol{x}}_o - \boldsymbol{x}\|_2. \tag{F.11}$$

We will use the following notations for the rest of the section

$$\widetilde{\boldsymbol{\Sigma}}_o = \boldsymbol{\Sigma}(\widetilde{\boldsymbol{x}}_o), \quad \widetilde{X}_o = \mathrm{diag}(\widetilde{\boldsymbol{x}}_o), \quad \widetilde{\boldsymbol{\Sigma}}_o = \boldsymbol{\Sigma}(\widetilde{\boldsymbol{x}}), \quad \widetilde{X} = \mathrm{diag}(\widetilde{\boldsymbol{x}}), \quad \widetilde{\boldsymbol{x}} \in N_{\delta_{\text{net}}}(\mathcal{C}_k). \tag{F.12}$$

Then in view of triangle inequality we get

$$\begin{aligned}
&\left|\vec{\boldsymbol{y}}^\top(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o))\right|\\
&\leq \left|\vec{\boldsymbol{y}}^\top(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o))\right| + \left|\vec{\boldsymbol{y}}^\top(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \mathrm{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o))\right|.
\end{aligned} \tag{F.13}$$

We use an union bound argument to control the first term above, uniformly over all choices of $\widetilde{\boldsymbol{x}} \in N_{\delta_{\text{net}}}(\mathcal{C}_k)$. This is done in the following result, which is the fixed A version of Definition 7.1.

Lemma F.1. Consider the definitions in (F.4) and (F.12). There exist constants $c_1, c_2, c_3, c_4 > 0$ such that the following holds with probability $1 - e^{-cn} - e^{-c_1 Lk \log((x_{\max}-x_{\min})n/\delta_{\text{net}})}$

$$\left| \vec{\boldsymbol{y}}^\top (\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\vec{\boldsymbol{y}} - \operatorname{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)) \right| \leq b_1 \sqrt{\mathscr{X}} + b_1', \quad \text{for all } \widetilde{\boldsymbol{x}}_o \in N_{\delta_{\text{net}}}(\mathcal{C}_k),$$

where, with the notation in (F.10), $b_1, b_1', \mathscr{X}$ are defined as

$$b_1 = c_3 \sqrt{k \log\left(\frac{(x_{\max} - x_{\min})n}{\delta_{\text{net}}}\right)}, \quad b_1' = c_4 C_{n,m,\sigma_z} k \log\left(\frac{(x_{\max} - x_{\min})n}{\delta_{\text{net}}}\right) \cdot x_{\max}^2, \qquad \text{(F.14)}$$
$$\mathscr{X} = \operatorname{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)).$$

The following result, a counterpart to Definition 7.1 for fixed A, controls the final term of (F.13).

Lemma F.2. Let C_{n,m,σ_z} be as in (F.10) and denote $b_2 = (C_{n,m,\sigma_z})^2 mL\delta_{\text{net}}$. There exist constants $c_1, c_2 > 0$ such that the following holds with probability $1 - e^{-c_1 n} - e^{-c_2 mL}$

$$\left| \vec{\boldsymbol{y}}^\top (\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \operatorname{Tr}(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \widehat{\boldsymbol{\Sigma}}_o)) \right| \leq b_2.$$

Combining Lemma F.2 with Lemma F.1, in view of (F.13) we have

$$\left| \vec{\boldsymbol{y}}^\top (\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)\vec{\boldsymbol{y}} - \operatorname{Tr}(\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\Sigma}_o - \widehat{\boldsymbol{\Sigma}}_o)) \right| \leq b_1 \sqrt{\mathscr{X}} + b_1' + b_2 \qquad \text{(F.15)}$$

Establishing a lower bound on $\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o)$: To find the lower bound, we use the decomposition

$$\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o) = \overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\widetilde{\boldsymbol{x}}_o) + \overline{f}(\widetilde{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o), \qquad \text{(F.16)}$$

with $\widetilde{\boldsymbol{x}}_o$ as in (F.11). The first term, $\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\widetilde{\boldsymbol{x}}_o)$ can be bounded by $C_{n,m,\sigma_z} x_{\max} n \delta_{\text{net}}$ using the fact that $\tilde{\boldsymbol{x}}_o$ is chosen to be at most δ_{net} distance away from $\hat{\boldsymbol{x}}_o$. We bound $\overline{f}(\widetilde{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o)$ using the following result, which is the fixed A version of Definition 7.3.

Lemma F.3. Assume that $\sigma_z^2 I_m + A\widetilde{X}_o^2 A^\top$ and $\sigma_z^2 I_m + AX_o^2 A^\top$, are invertible. Then,

$$\overline{f}(\widetilde{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o) \geq \frac{1}{2(1+\widetilde{\lambda}_{\max})^2} \operatorname{Tr}\left(\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\boldsymbol{\Sigma}_o^{-1}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\right), \qquad \text{(F.17)}$$

where $\widetilde{\lambda}_{\max} > 0$ is the maximum singular value of $\boldsymbol{\Sigma}_o^{-\frac{1}{2}}(\widetilde{\boldsymbol{\Sigma}}_o - \boldsymbol{\Sigma}_o)\boldsymbol{\Sigma}_o^{-\frac{1}{2}}$. Moreover, $\widetilde{\lambda}_{\max} \leq C_{n,m,\sigma_z}$ on the event $\mathcal{S}_{\text{fixed}}$ in (F.2).

The following result, a fixed A version of Definition 7.4, controls $|\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\widetilde{\boldsymbol{x}}_o)|$ for a given δ_{net}.

Lemma F.4. $|\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\widetilde{\boldsymbol{x}}_o)| \leq C_{n,m,\sigma_z} \cdot x_{\max} n \delta_{\text{net}} \ll 1$ with probability $1 - e^{-cn}$ for some $c > 0$.

Combining the above results, in view of (F.16) we have, with probability $1 - \exp(-cn)$,

$$\overline{f}(\widehat{\boldsymbol{x}}_o) - \overline{f}(\boldsymbol{x}_o) \geq \frac{\mathscr{X}}{(C_{n,m,\sigma_z})^2} - 1. \qquad \text{(F.18)}$$

Simplifying the quadratic inequality: Combining (F.18) and (F.15), in view of (F.9), we have

$$\mathbb{P}\left[\frac{\mathscr{Z}}{(C_{n,m,\sigma_z})^2} \le b_1\sqrt{\mathscr{Z}} + b_1' + b_2 + 1\right] \ge 1 - e^{-c_1 Lk\log((x_{\max}-x_{\min})n/\delta_{\text{net}})} - e^{-cn} - \exp\left(-cmL\right). \tag{F.19}$$

Rewrite the last inequality as $az^2 - bz - c \le 0$, with $z = \sqrt{\mathscr{Z}}, a = \frac{1}{(C_{n,m,\sigma_z})^2}, b = b_1, c = b_1' + b_2 + 1$. As $z = \sqrt{\mathscr{Z}} > 0$, z^2 is smaller than the square of the positive root of $az^2 - bz - c = 0$, which implies

$$\mathscr{Z} = z^2 \le \left(\frac{-b + \sqrt{b^2 + 4ac}}{2a}\right)^2 \le \left(\frac{-b + \sqrt{b^2 + 4ac}}{2a}\right)\left(\frac{b + \sqrt{b^2 + 4ac}}{2a}\right) = \frac{c}{a}, \tag{F.20}$$

where the second inequality followed as $a, b, c > 0$. Using the notations from, (F.10), Definition F.1 and Definition F.2 we get

$$\mathscr{Z} \le \frac{b_1' + b_2}{a} = (C_{n,m,\sigma_z})^2\left(c_3 k \log\left(\frac{(x_{\max} - x_{\min})n}{\delta_{\text{net}}}\right) \cdot x_{\max}^2 + (C_{n,m,\sigma_z})^2 mL\delta_{\text{net}}\right). \tag{F.21}$$

Choose $\delta_{\text{net}} = \frac{x_{\max}}{n^5}$ and recall $mL \le n^4 k \log n$ from Definition 2.14. Then, from the last display we use (F.19) to get for a constant $C > 0$

$$\mathbb{P}\left[\mathscr{Z} \le C \cdot (C_{n,m,\sigma_z})^2 k \log n\right] = 1 - O\left(n^{-ckL} + \exp(-cn) + 2\exp\left(-cmL\right)\right). \tag{F.22}$$

Finding a lowerbound for $\mathscr{Z}$: In view of Definition 4.14, using the block structure of $\mathbf{\Sigma}_o, \widetilde{\mathbf{\Sigma}}_o$ given in (F.4), we have on the event $\mathcal{E}_{\text{sing}}$,

$$\begin{aligned}
\mathscr{Z} &= \operatorname{Tr}\left[\mathbf{\Sigma}_o^{-1}\left(\widetilde{\mathbf{\Sigma}}_o - \mathbf{\Sigma}_o\right)\mathbf{\Sigma}_o^{-1}\left(\widetilde{\mathbf{\Sigma}}_o - \mathbf{\Sigma}_o\right)\right] \\
&= L\operatorname{Tr}\left[(\Sigma(\boldsymbol{x}_o)^{-1}(\Sigma(\widetilde{\boldsymbol{x}}_o) - \Sigma(\boldsymbol{x}_o))\Sigma(\boldsymbol{x}_o)^{-1}(\Sigma(\widetilde{\boldsymbol{x}}_o) - \Sigma(\boldsymbol{x}_o))\right] \\
&\ge \frac{L}{(C_{n,m,\sigma_z})^2\left(\sigma_z^2 + x_{\max}^2 E_{\max}\right)^2}\left\|A(\widetilde{X}_o^2 - X_o^2)A^\top\right\|_{\text{HS}}^2,
\end{aligned} \tag{F.23}$$

where C_{n,m,σ_z} is as in (F.10). The lower bound on $\mathscr{Z}$ is completed with the following lower bound on $\left\|A(\widetilde{X}_o^2 - X_o^2)A^\top\right\|_{\text{HS}}^2$. The result is the fixed A and $L = 1$ version of Definition 7.5.

Lemma F.5. The following holds true with a probability $1 - \exp\left(-2k\log n\right) - m\exp\left(-cn\right)$

$$\begin{aligned}
&\|A(\widetilde{X}_o^2 - X_o^2)A^\top\|_{\text{HS}}^2 \\
\ge& 4m(m-1)x_{\min}^2\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2 - 4Cx_{\max}^2\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2\log m\sqrt{mn}\sqrt{k\log n} - Cx_{\max}^4 nk\log m\log n.
\end{aligned}$$

Final upper bound on $\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2$: We combine (F.23), (F.22), and Definition F.5 to summarize the above in terms of the following quadratic inequality with respect to $\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2$, that holds with a probability $1 - O\left(n^{-ckL} + \exp(-cn) + \exp\left(-cmL\right)\right)$

$$
\begin{aligned}
& a\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2 - b\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2 - d \le 0 \\
& a = \frac{C_1 m(m-1)x_{\min}^2}{(C_{n,m,\sigma_z})^2(\sigma_z^2 + x_{\max}^2 E_{\max})^2}, \quad b = \frac{C_2 x_{\max}^2 n \log m \sqrt{mk\log n}}{(C_{n,m,\sigma_z})^2(\sigma_z^2 + x_{\max}^2 E_{\max})^2}, \\
& d = \frac{C_3 x_{\max}^4 nk \log m \log n}{C \cdot (C_{n,m,\sigma_z})^2(\sigma_z^2 + x_{\max}^2 E_{\max})^2} + C\cdot (C_{n,m,\sigma_z})^2 \frac{k\log n}{L}.
\end{aligned} \tag{F.24}
$$

In view of an argument similar to (7.22) we have with a probability $1-O\left(n^{-ckL} + \exp(-cn) + \exp\left(-cmL\right)\right)$

$$
\frac{1}{n}\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2 \le \frac{d}{na} \le \frac{C_3 x_{\max}^4}{x_{\min}^2}\frac{k\log m \log n}{m^2} + \frac{(C_{n,m,\sigma_z})^4(\sigma_z^2 + x_{\max}^2 E_{\max})^2 k\log n}{nm^2 L}. \tag{F.25}
$$

This implies, in view of $\frac{1}{n}\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2 \le x_{\max}^2$,

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\|\widetilde{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2\right] \le & \frac{C_3 x_{\max}^4}{x_{\min}^2}\frac{k\log m \log n}{m^2} + \frac{(C_{n,m,\sigma_z})^4(\sigma_z^2 + x_{\max}^2 E_{\max})^2 k\log n}{nm^2 L} \\
& + C_1 x_{\max}^2 (n^{-ckL} + L\exp(-cn) + \exp\left(-cmL\right)).
\end{aligned} \tag{F.26}
$$

As $\|\widetilde{\boldsymbol{x}}_o - \widehat{\boldsymbol{x}}_o\|_2 \le \delta_{\text{net}} \le \frac{x_{\max}}{n^5}$ from the definition in (7.13), we continue the last display to get

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\|\widehat{\boldsymbol{x}}_o - \boldsymbol{x}_o\|_2^2\right] \le 2C_4 \Bigg\{ & \frac{x_{\max}^4}{x_{\min}^2}\frac{k\log m \log n}{m^2} + \frac{(C_{n,m,\sigma_z})^4(\sigma_z^2 + x_{\max}^2 E_{\max})^2 k\log n}{nm^2 L} \\
& + x_{\max}^2 (n^{-ckL} + L\exp(-cn) + \exp\left(-cmL\right)) + \frac{x_{\max}^2}{n^{10}} \Bigg\}.
\end{aligned} \tag{F.27}
$$

Note by our assumption $\log m \ll n$. Therefore the first term has a slower growth rate compared to the second term.

To summarize, we have in the regime $n \ge 4m$, or in the regime $n < 4m$ but $\sigma_z^2 \ge m$ that

$$
R_2^{\dagger}(\mathcal{C}, m, n, \sigma_z^2) = O_{x_{\min}, x_{\min}}\left(\frac{\max(\sigma_z^4, n^2)}{m^2 n}\frac{k\log n}{L} + \frac{k\log m \log n}{m^2}\right). \tag{F.28}
$$

For the case $n < 4m, \sigma_z^2 \le m$, we have by monotonicity of the risk in σ_Z (Definition 4.2) that

$$
R_2^{\dagger}(\mathcal{C}, m, n, \sigma_z^2) \le R_2^{\dagger}(\mathcal{C}, m, n, m) = O_{x_{\min}, x_{\min}}\left(\frac{\max(m^2, n^2)}{m^2 n}\frac{k\log n}{L} + \frac{k\log m \log n}{m^2}\right). \tag{F.29}
$$

Combining these, we have

$$
R_2^{\dagger}(\mathcal{C}, m, n, \sigma_z^2) = O_{x_{\min}, x_{\min}}\left(\frac{\max(\sigma_z^4, m^2, n^2)}{m^2 n}\frac{k\log n}{L} + \frac{k\log m \log n}{m^2}\right). \tag{F.30}
$$

When $\frac{\max(\sigma_z^4, m^2, n^2)}{m^2 n}\frac{k\log n}{L} \ge \frac{k\log m \log n}{m^2}$, or equivalently $\max(\sigma_z^4, m^2, n^2) \ge nL\log m$, we have

$$
R_2^{\dagger}(\mathcal{C}, m, n, \sigma_z^2) = O_{x_{\min}, x_{\min}}\left(\frac{\max(\sigma_z^4, n^2)}{m^2 n}\frac{k\log n}{L}\right). \tag{F.31}
$$

F.2 Proof of Definition 2.15 in the case $n \geq 4m$

The proof follows the same approach as in Section 5 and the only difference here is the argument in bounding the KL divergence and β_r. We shall point out the difference in below.

Let $A = A_1 = \cdots = A_L$ be an $m \times n$ Gaussian matrix, and let $\mathcal{X}_k, \mathcal{X}^{\text{finite}}, \mathcal{S}_{\text{sep}}, r, N_{\text{div}}, \alpha_r$ and δ_r be chosen as in Section 5. In place of Definition 5.2, we have (with the same proof)

Lemma F.6. Denote $E_{\max} := \lambda_{\max}(AA^T), E_{\min} := \lambda_{\min}(AA^T)$. On the event $\mathcal{E}_{\text{sing}}$, defined in (4.8), if $\frac{[\sigma_z^2 + x_{\max}^2 \cdot E_{\max}] \cdot E_{\max}}{(\sigma_z^2 + x_{\min}^2 \cdot E_{\min})^2} \cdot x_{\max} \delta_r < \frac{1}{4}$, we have for all $\boldsymbol{x}_i \neq \boldsymbol{x}_j \in \mathcal{S}_{\text{sep}}$

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}_i} \parallel \mathbb{P}_{\boldsymbol{x}_j}) \leq 2 \frac{(\sigma_z^2 + x_{\max}^2 E_{\max})^2}{(\sigma_z^2 + x_{\min}^2 E_{\min})^4} \cdot L \left\| A(X_i^2 - X_j^2)A^\top \right\|_{\text{HS}}^2,$$

where X_i and X_j are diagonal matrices corresponding to the vectors $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{S}_{\text{sep}}$.

We now apply the upper tail bound of Definition 4.10 for the case when $L = 1$ to find a deterministic upper bound for $L \left\| A(X_i^2 - X_j^2)A^\top \right\|_{\text{HS}}^2$. We set $\boldsymbol{d}_{i,j} := \boldsymbol{x}_i^2 - \boldsymbol{x}_j^2$, and define $D_{i,j} = \mathrm{diag}(\boldsymbol{d}_{i,j})$. Choose We choose the following values of $t_{1,i,j}$ and $t_{2,i,j}$ to apply the upper tail bound in Definition 4.10

$$\begin{aligned}
t_{1,i,j} :=& C_{t_1} \left(x_{\max} \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|_2 \sqrt{\log(mr^2)} + x_{\max} \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|_\infty \log(mr^2) \right); \\
t_{2,i,j} :=& C_{t_2} \log m \left(x_{\max}^2 \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|_4^2 \sqrt{m(\sqrt{m} + \sqrt{n})^4 \log r^2} + x_{\max}^2 \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|_\infty^2 (\sqrt{n} + \sqrt{m})^2 \log r^2 \right),
\end{aligned}$$

where C_{t_1} and C_{t_2} are two constants. In view of the above definition, consider the event

$$\mathcal{E}_{\text{dcpl}} := \bigcap_{1 \leq i < j \leq r} \left[\| A(X_i^2 - X_j^2)A^\top \|_{\text{HS}}^2 < m t_{1,i,j}^2 + m(m-1) \| \boldsymbol{d}_{i,j} \|_2^2 + t_{2,i,j} \right]. \tag{F.32}$$

Using the same argument as in Subsection 5.4, we have

$$\mathbb{P}(\mathcal{E}_{\text{dcpl}}^c \cap \widetilde{\mathcal{E}}_{\text{maxsing}}) \leq r^2 \exp \left\{ -\tilde{C} \left(\log(mr^2) + (\log m)(\log r) \right) \right\}, \tag{F.33}$$

and for sufficiently large C_{t_1}, C_{t_2}, we have

$$\mathbb{P}(\mathcal{E}_{\text{dcpl}} \cap \widetilde{\mathcal{E}}_{\text{maxsing}}) \geq 1 - \frac{1}{(rm)^8} - 2m \exp(-cn). \tag{F.34}$$

In view of the above, on the high-probability event $\mathcal{E}_{\text{dcpl}} \cap \widetilde{\mathcal{E}}_{\text{maxsing}}$, we have for each $1 \leq i < j \leq r$,

$$\begin{aligned}
\left\| A_l(X_i^2 - X_j^2)A_l^\top \right\|_{\text{HS}}^2 \leq& m t_{1,i,j}^2 + m(m-1) \| \boldsymbol{d}_{i,j} \|_2^2 + t_{2,i,j} \\
&+ C \log m \left(x_{\max}^2 \sqrt{\frac{kn}{N_{\text{div}}}} \delta_r^2 \sqrt{m(\sqrt{m} + \sqrt{n})^4 \log r^2} + \delta_r^2 (\sqrt{n} + \sqrt{m})^2 \log r^2 \right),
\end{aligned}$$

Hence, restricting to the event $\mathcal{E}_{\mathrm{dcpl}} \cap \widetilde{\mathcal{E}}_{\mathrm{maxsing}} \cap \mathcal{E}_{\mathrm{sing}}$, together with Definition 5.2, we have for constant $\bar{C} := C_{x_{\min},x_{\max}} > 0$

$$
\begin{aligned}
&\beta_r := \max_{1\le i<j\le r} \mathrm{KL}(\mathbb{P}_i \,\|\, \mathbb{P}_j) \\
&\le 2\frac{\left(\sigma_z^2 + x_{\max}^2 \cdot E_{\max}\right)^2 L}{\left(\sigma_z^2 + x_{\min}^2 \cdot \frac{1}{4}(\sqrt{n}-\sqrt{m})^2\right)^4} \max_{1\le i<j\le r} \left\| A(X_i^2 - X_j^2)A^\top \right\|_{\mathrm{HS}}^2 \\
&\le \frac{\bar{C}L}{\max(\sigma_z^4, n^2)} \Bigg(m\frac{kn}{N_{\mathrm{div}}}\delta_r^2 \log(mr^2) + m\delta_r^2 \log^2(mr^2) + m(m-1)\frac{kn}{N_{\mathrm{div}}}\delta_r^2 \\
&\qquad + \log m \sqrt{\frac{kn}{N_{\mathrm{div}}}}\delta_r^2 \sqrt{m(\sqrt{m}+\sqrt{n})^4 \log r^2} + \delta_r^2(\sqrt{n}+\sqrt{m})^2(\log m)(\log r^2) \Bigg) \\
&\le \frac{\bar{C}\delta_r^2 m^2 nLk}{\max(\sigma_z^4, n^2)N_{\mathrm{div}}} \left(\frac{\log(mr^2)}{m} + \frac{\log^2(mr^2)N_{\mathrm{div}}}{mnk} + 1 + \sqrt{\frac{nN_{\mathrm{div}}\log r^2}{km^3/(\log m)^2}} + \frac{(\log m)(\log r^2)N_{\mathrm{div}}}{m^2} \right) \\
&\le \Theta_{x_{\min},x_{\max}}(1) \frac{m^2 nLk}{\max(\sigma_z^4, n^2)N_{\mathrm{div}}}\delta_r^2,
\end{aligned}
\tag{F.35}
$$

where the last inequality followed by factoring out δ_r^2 and using the following inequalities that are consequences of Definition 5.1, alongside our assumptions $\log m = \Theta(\log n), \log L = O(\log n)$, and there exists $\varepsilon \in (0, 1/2)$ such that $k \le n^{1-2\varepsilon}, \max(\sigma_z^4, m^2, n^2)k \log n \le m^2 n^{1-\varepsilon} L$. Note that the bound (F.35) is the same as the bound (5.19) for β_r.

As a consequence, by Definition 4.3 we have for any estimator $\widehat{\boldsymbol{x}}$, the same lower bound

$$
\begin{aligned}
\max_{1\le i\le r} \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_i\|^2}{n}\right] &\ge \frac{\alpha_r^2}{4n}\left(1 - \frac{\beta_r + \log 2}{\log r}\right)^2 \mathbb{P}\left(\widetilde{\mathcal{E}}_{\mathrm{maxsing}} \cap \mathcal{E}_{\mathrm{dcpl}} \cap \mathcal{E}_{\mathrm{sing}}\right) = \Theta\left(\frac{\alpha_r^2}{n}\right) \\
&= \Theta_{x_{\max},x_{\min}}\left(\frac{\max(\sigma_z^4, n^2)}{m^2 n}\frac{k\log\left(N_{\mathrm{div}}/k\right)}{L}\right) = \Theta_{\varepsilon,x_{\max},x_{\min}}\left(\frac{\max(\sigma_z^4, n^2)k\log n}{m^2 nL}\right).
\end{aligned}
$$

F.3 Proof of Definition 2.15 in the case $n \le 4m$

The proof follows the same approach as in Section 6 and here we point out the main differences. This section will primarily establish the lower bound for the sub-case $m \ge 4n, \sigma_z^2 = 0$ given by

$$
R_2^\dagger(\mathcal{C}_k, m, n, 0) = \inf_{\boldsymbol{\delta}} \sup_{\boldsymbol{x}\in\mathcal{C}_k} \mathbb{E}\left[\frac{\left\|\boldsymbol{\delta}(\vec{\boldsymbol{y}}) - \boldsymbol{x}_o\right\|_2^2}{n}\right] = \Omega_{\varepsilon,x_{\min}x_{\max}}\left(\frac{k\log n}{nL}\right), \quad m \ge 4n. \tag{F.36}
$$

Then, the lower bound for a general $\sigma_z^2 \ge 0$ and $4n \ge m \ge \frac{n}{4}$ follow from Definition 4.1 and Definition 4.2 by the same monotonicity argument as in Section 6 before (6.2). Note that, in view of Definition 4.1, for any $\frac{n}{4} \le m \le 4n$, the last display implies

$$
R_2^\dagger(\mathcal{C}_k, m, n, 0) \ge R_2^\dagger(\mathcal{C}_k, n/4, n, 0) \ge R_2^\dagger(\mathcal{C}_k, 4n, n, 0) = \Omega_{\varepsilon,x_{\min},x_{\max}}\left(\frac{k\log n}{nL}\right). \tag{F.37}
$$

By the same argument as in the paragraph before (6.6), in the case $4m \geq n$ and $\sigma_z^2 \geq m$ we have the lower bound

$$R_2^{\dagger}(\mathcal{C}_k, m, n, \sigma_z) = \Omega_{x_{\max}, x_{\min}}\left(\frac{\sigma_z^4}{m^2 n} \cdot \frac{k \log n}{L}\right), \text{ whenever } \sigma_z^2 \geq m. \tag{F.38}$$

To achieve a lower bound for the sub-case $m \geq \frac{n}{4}, \sigma_z^2 \leq m$, we first use that the miminax error is non-decreasing function in σ_z (Definition 4.2) to get $R_2^{\dagger}(\mathcal{C}_k, m, n, 0) \leq R_2^{\dagger}(\mathcal{C}_k, m, n, \sigma_z)$. Then, to achieve a lower bound to $R_2^{\dagger}(\mathcal{C}_k, m, n, 0)$ we combine the lower bounds in (F.36) and (F.37) to get

$$R_2(\mathcal{C}_k, m, n, 0) = \Omega_{x_{\max}, x_{\min}}\left(\frac{m^2}{m^2 n} \cdot \frac{k \log n}{L}\right) = \Omega_{x_{\max}, x_{\min}}\left(\frac{k \log n}{nL}\right). \tag{F.39}$$

Then, for $\sigma_z^2 \leq m$, by (6.3) we get

$$C_1 \frac{k \log n}{nL} \leq R_2^{\dagger}(\mathcal{C}_k, m, n, 0) \leq R_2^{\dagger}(\mathcal{C}_k, m, n, \sigma_z), \tag{F.40}$$

where C_1 is a constant depending on $x_{\min}, x_{\max}$. Combining (F.38) and (F.39) for the case $n \leq 4m$ yields the desired minimax lower bound

$$R_2^{\dagger}(\mathcal{C}_k, m, n, \sigma_z) = \Omega_{x_{\max}, x_{\min}}\left(\frac{\max(\sigma_z^4, m^2)}{m^2 n} \cdot \frac{k \log n}{L}\right), \quad m \geq \frac{n}{4}, \sigma_z \geq 0.$$

Now it remains to establish (F.36). We will show this with sufficient statistics as in Section 6. Define $\boldsymbol{A} = \mathrm{diag}(A, \dots, A) \in \mathbb{R}^{mL \times nL}$ and note that $\boldsymbol{A}^\top \boldsymbol{A} = \mathrm{diag}(A^\top A, \dots, A^\top A)$. Throughout the section we analyze the expected loss on the high probability event $\mathcal{E}'_{\text{sing}}$ defined in (4.10), where $A = A_1 = \cdots = A_L$ and $A^\top A$ is invertible. Then with the same proof as Definition 6.2, we have

Proposition F.7. Consider the case $\sigma_z = 0$ and that the event $\mathcal{E}'_{\text{sing}}$ holds. Then $\boldsymbol{T_A}(\vec{\boldsymbol{y}}) = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \vec{\boldsymbol{y}}$ is a sufficient statistic for the parameter $\boldsymbol{x}_o$.

By Rao-Blackwell theorem (Definition 6.3), we have,

$$\mathbb{E}\left[\left\|\boldsymbol{\delta}(\vec{\boldsymbol{y}}) - \boldsymbol{x}_o\right\|_2^2 | \mathcal{E}'_{\text{sing}}\right] \geq \mathbb{E}\left[\left\|g(\boldsymbol{T_A}(\vec{\boldsymbol{y}})) - \boldsymbol{x}_o\right\|_2^2 | \mathcal{E}'_{\text{sing}}\right].$$

Therefore it suffices to prove the following lemma

Lemma F.8. Consider the model (1.2) with $\sigma_z = 0, m \geq 4n$. Then, there exists a constant $C \geq 0$, we have

$$\inf_g \sup_{\boldsymbol{x} \in \mathcal{C}_k} \mathbb{E}\left[\left\|g(\boldsymbol{T_A}(\vec{\boldsymbol{y}})) - \boldsymbol{x}_o\right\|_2^2 | \mathcal{E}'_{\text{sing}}\right] \geq C \frac{k \log n}{L}.$$

The proof of Definition F.8 is the same as that of Definition 6.4. Finally, it follows from it follows from Theorem 6.3 and Definition F.8 that

$$\inf_{\boldsymbol{\delta}} \sup_{\boldsymbol{x} \in \mathcal{C}_k} \mathbb{E}\left[\frac{\left\|\boldsymbol{\delta}(\vec{\boldsymbol{y}}) - \boldsymbol{x}_o\right\|_2^2}{n}\right] \geq \inf_{\boldsymbol{\delta}} \sup_{\boldsymbol{x} \in \mathcal{C}_k} \mathbb{E}\left[\left.\frac{\left\|g(\boldsymbol{T_A}(\vec{\boldsymbol{y}})) - \boldsymbol{x}_o\right\|_2^2}{n}\right| \mathcal{E}'_{\text{sing}}\right] \mathbb{P}\left(\mathcal{E}'_{\text{sing}}\right) = \Omega\left(\frac{k \log n}{nL}\right).$$

G Proof of Theorem 2.10

G.1 Proof of the lower bound

For

$$\mathcal{S}_k^{\text{bdd}} := \left\{ \boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_o \leq k, x_i = 0 \text{ or } 0 < x_{\min} \leq x_i \leq x_{\max} \right\},$$

and any $S \subset [n]$ and $|S| \leq k$,

$$\mathcal{C}_S := \left\{ \boldsymbol{x}_S \in \mathbb{R}^{|S|} : x_{\min} \leq x_i \leq x_{\max}, i \in S \right\} = [x_{\min}, x_{\max}]^{|S|}.$$

Also, for a fixed $S \subset [n]$ with $|S| \leq k$, define

$$\mathcal{S}_{k,S}^{\text{bdd}} = \{ \boldsymbol{x} \in \mathbb{R}^n : x_i = 0, i \notin S, \text{ and } x_i \in [x_{\min}, x_{\max}] \text{ for } i \in S \}.$$

In this section, we aim to prove the lower bound, i.e we aim to prove that

$$\begin{aligned} & R_2(\mathcal{S}_k^{\text{bdd}}, m, n, k, L, \sigma_z) \\ & = \inf_{\widehat{\boldsymbol{x}} \in \mathbb{R}^n} \sup_{\boldsymbol{x}_o \in \mathcal{S}_k^{\text{bdd}}} \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2}{n} \right] = \Omega_{x_{\max}, x_{\min}}\left(\frac{k}{nL} \right). \end{aligned} \tag{G.1}$$

First note that by the monotonicity result proved in definition 4.2, we have $R_2(\mathcal{S}_k^{\text{bdd}}, m, n, k, L, \sigma_z) \geq R_2(\mathcal{S}_k^{\text{bdd}}, m, n, k, L, 0)$. Therefore it suffices to establish

$$R_2(\mathcal{S}_k^{\text{bdd}}, m, n, k, L, 0) = \Omega_{x_{\max}, x_{\min}}\left(\frac{k}{nL} \right). \tag{G.2}$$

For any $S \subset [n]$ and $|S| \leq k$, since $\mathcal{S}_{k,S}^{\text{bdd}} \subset \mathcal{S}_k^{\text{bdd}}$

$$\inf_{\widehat{\boldsymbol{x}} \in \mathbb{R}^n} \sup_{\boldsymbol{x}_o \in \mathcal{S}_k^{\text{bdd}}} \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2}{n} \right] \geq \inf_{\widehat{\boldsymbol{x}} \in \mathbb{R}^n} \sup_{\boldsymbol{x}_o \in \mathcal{S}_{k,S}^{\text{bdd}}} \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2}{n} \right] \inf_{\widehat{\boldsymbol{x}}_S \in \mathbb{R}^{|S|}} \sup_{\boldsymbol{x}_{o,S} \in \mathcal{C}_S} \mathbb{E}\left[\frac{\left\|\widehat{\boldsymbol{x}}_S - \boldsymbol{x}_{o,S}\right\|_2^2}{n} \right].$$

To obtain the last equality we have noted that since we know the exact location of non-zero elements of $\boldsymbol{x}_o$ for any $\boldsymbol{x}_o \in \mathcal{S}_{k,S}^{\text{bdd}}$, we have set the value of $\hat{\boldsymbol{x}}$ to zero at those locations, and have reduced the problem to estimating the nonzero elements of $\boldsymbol{x}_o$. We now claim that, in particular, if $|S| = k$, we have

$$R_2(\mathcal{C}_S, m, n, k, L, 0) = \Omega_{x_{\max}, x_{\min}}\left(\frac{k}{nL} \right). \tag{G.3}$$

Indeed, in this case our model reduces to

$$\boldsymbol{y}_l = A_{l,S} X_{o,S} \boldsymbol{w}_{l,S}, \quad l = 1, 2, ..., L. \tag{G.4}$$

where $A_{l,S}$ is the $m \times k$ matrix of whose columns are those of A_l with indices in S, $X_{o,S}$ is the $k \times k$ diagonal matrix in which all the diagonal locations with indices in S- are nonzero, and $\boldsymbol{w}_{l,S}$ is the k-dimensional Gaussian vector. Since $m \gg k$, we know that $\widetilde{\boldsymbol{y}}_l := (A_{l,S}^\top A_{l,S})^{-1} A_{l,S}^\top \boldsymbol{y}_l$ is sufficient statistics (see definition 6.2) and by using Rao-Blackwell theorem (Theorem 6.3), it suffices to obtain a lower bound for the minimax risk of estimation $\boldsymbol{x}_S$ from $\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2, \ldots, \widetilde{\boldsymbol{y}}_L$:

$$\widetilde{\boldsymbol{y}}_l := X_{o,S}\boldsymbol{w}_{l,S}, \quad l = 1, 2, ..., L. \tag{G.5}$$

We shall prove this by Definition 4.3. To use Definition 4.3, note that the joint distribution of $\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2, \ldots, \widetilde{\boldsymbol{y}}_L$ is given by:

$$\mathbb{P}_{\boldsymbol{x}} \sim \otimes_{l=1}^{L} N(\mathbf{0}, X_{o,S}^2). \tag{G.6}$$

For $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$, we define the pseudo-metric,

$$d(\theta(\mathbb{P}_{\boldsymbol{x}}), \theta(\mathbb{P}_{\boldsymbol{x}'})) = d(\boldsymbol{x}, \boldsymbol{x}') := \|\boldsymbol{x} - \boldsymbol{x}'\|_2. \tag{G.7}$$

Next, we shall construct an α_r-separated set $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_r\}$ in $[x_{\min}, x_{\max}]^{|S|} = [x_{\min}, x_{\max}]^k$. To construct this subset, we use the following steps. Using the δ-packing defined in definition 4.5 we have:

Lemma G.1. For any $p < \frac{k}{2}$, the k-dimensional Hamming cube $\{0, 1\}^k$ has p-packing number at least $\frac{2^k}{\sum_{i=0}^{p} \binom{k}{i}}$ with respect to the Hamming distance.

Proof. Let $\mathcal{S}$ be a p-packing subset of $\{0, 1\}^k$. By definition we must have

$$\{0, 1\}^k \subset \cup_{\boldsymbol{x} \in \mathcal{S}} B(\boldsymbol{x}, p), \tag{G.8}$$

where $B(\boldsymbol{x}, p)$ is the ball centered at $\boldsymbol{x}$ with radius p with respect to the Hamming distance. In other words, $B(\boldsymbol{x}, p)$ is the collection of points in $\{0, 1\}^k$ whose at most p coordinates are different from those of $\boldsymbol{x}$. A direct counting gives $|B(\boldsymbol{x}, p)| \leq \sum_{i=0}^{p} \binom{k}{i}$. Taking cardinality of both sides of (G.8) yields

$$2^k \leq \sum_{i=0}^{p} \binom{k}{i} |\mathcal{S}|.$$

This completes the proof. □

To obtain a simpler lower bound for the p-packing that can be used in our arguments, we assume that $p = k/7$ and establish the following lemma:

Lemma G.2. We have $\frac{2^k}{\sum_{i=0}^{\lfloor k/7 \rfloor} \binom{k}{i}} \geq a^k$ for some $a > 1$.

Proof. Indeed, for $p := \lfloor k/7 \rfloor \leq \frac{k}{2}$, since $\binom{k}{i-1} / \binom{k}{i} = \frac{i}{k-i+1} \leq \frac{1}{2}$, for $i \leq p \leq k/2$, we have

$$\sum_{i=0}^{p} \binom{k}{i} \leq \left(1 + \frac{1}{2} + \cdots \frac{1}{2^{p-1}}\right) \binom{k}{p} \leq 2 \binom{k}{p} \leq 2 \left(\frac{ek}{p}\right)^p.$$

So $\log \frac{2^k}{\sum_{i=0}^{p} \binom{k}{i}} \geq k \log 2 - p \log \left(ek/p\right) - \log 2 \gg k \cdot b$ for some $b > 0$. Selecting $a = e^b$ establishes the result. □

Let $\mathcal{H} \subset \{0,1\}^k$ be an $\lfloor k/7 \rfloor$-packing set with respect to the Hamming distance from Definition G.1. By Definition G.2, $|\mathcal{H}| \geq a^k$.

Now we aim to use $\mathcal{H}$ to create a subset of $\mathcal{C}_S$. Define $\bar{x} := \frac{x_{\min}+x_{\max}}{2}$. We create the subset by scaling and translation of $\mathcal{H}$ in the following way:

$$c\frac{x_{\max}-x_{\min}}{\sqrt{L}} \cdot \mathcal{H} + \overline{\boldsymbol{x}} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_r\} \subset B_{\ell_2}\left(\overline{\boldsymbol{x}}, c\sqrt{\frac{k}{L}} \cdot (x_{\max}-x_{\min})\right) \cap [x_{\min}, x_{\max}]^k. \tag{G.9}$$

Note that this set forms a $c\frac{x_{\max}-x_{\min}}{\sqrt{L}}\sqrt{\lfloor k/7 \rfloor}$ -separated set for $[x_{\min}, x_{\max}]^k$ with respect to the ℓ_2-distance, with cardinality $r \geq a^k$ for some $a > 1$. To use Definition 4.3 set $\alpha_r := c\frac{x_{\max}-x_{\min}}{\sqrt{L}}\sqrt{\lfloor k/7 \rfloor}$. It follows from Definition 5.2 (with A_l replaced by I_n) that

$$\begin{aligned}
\beta_r := \max_{1\leq i<j\leq r} \mathrm{KL}(\mathbb{P}_i \parallel \mathbb{P}_j) &\leq L\frac{x_{\max}^4}{x_{\min}^8}\max_{1\leq i<j\leq r}\left\|X_i^2 - X_j^2\right\|_{\mathrm{HS}}^2 \\
\leq \frac{4x_{\max}^6}{x_{\min}^8}L\max_{1\leq i<j\leq r}\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|_2^2 &\leq \frac{4x_{\max}^6}{7x_{\min}^8}(x_{\max}-x_{\min})^2c^2k \leq \frac{1}{10}\log r,
\end{aligned}$$

for sufficiently small c from (G.9). Hence, by Definition 4.3,

$$\inf_{\widehat{\boldsymbol{x}}_S\in\mathbb{R}^{|S|}}\sup_{\boldsymbol{x}_{o,S}\in\mathcal{C}_S}\mathbb{E}\left[\frac{\left\|\widehat{\boldsymbol{x}}_S-\boldsymbol{x}_{o,S}\right\|_2^2}{n}\right] \geq \frac{\alpha_r^2}{4n}\left(1-\frac{\beta_r+\log 2}{\log r}\right)^2\mathbb{P}\left(\mathcal{E}'_{\mathrm{sing}}\right) = \Theta\left(\frac{\alpha_r^2}{n}\right) = \Theta_{x_{\min},x_{\max}}\left(\frac{k}{nL}\right).$$

G.2 Proof of the upper bound

In this section we aim to prove the upper bound. In other words, we aim to prove that

$$R_2(\mathcal{S}_k^{\mathrm{bdd}}, m, n, k, L, \sigma_z) = \inf_{\widehat{\boldsymbol{x}}\in\mathbb{R}^n}\sup_{\boldsymbol{x}_o\in\mathcal{S}_k^{\mathrm{bdd}}}\mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_o\|_2^2}{n}\right] = O_{x_{\max},x_{\min}}\left(\frac{k}{nL}+\frac{\sigma_z^2 k\log(n/k)}{mn}\right)$$

Consider the model

$$\boldsymbol{y}_l = A_l X_o \boldsymbol{w}_l + \boldsymbol{z}_l, \text{ for } l = 1, \ldots, L. \tag{G.10}$$

Define, $\boldsymbol{u}_l := X_o\boldsymbol{w}_l \sim N(0, X_o^2)$. Observe that $\boldsymbol{u}_l$ is unbounded, but still k-sparse. If we think of u_l as our new unkown data, the model (G.10) reduces to L copies of classical sparse linear regression models

$$\boldsymbol{y}_l = A_l\boldsymbol{u}_l + \boldsymbol{z}_l, \quad l = 1, \ldots, L,$$

where $\boldsymbol{u}_l \in \mathcal{S}_k := \{\boldsymbol{x} \in \mathbb{R}^k : \|\boldsymbol{x}\|_0 \leq k\}$. Our strategy for obtaining the upper bound is to first estimate each $\boldsymbol{u}_\ell$ separately from $\boldsymbol{y}_l$. This is a standard problem in sparse linear regression. We use one of the classic results for obtaining an upper bound for the minimax risk of sparse linear regression model:

Theorem G.3. (Verzelen, 2012, Equation (3.9) and Proposition 6.4) Let A be an $m \times n$ Gaussian matrix and $\boldsymbol{z} \sim N(0, \sigma_z^2 I_n)$. Consider the sparse linear regression model $\boldsymbol{y} = A\boldsymbol{u} + \boldsymbol{z}$ where the unknown signal $\boldsymbol{u} \in \mathcal{S}_k$. Then for $k\log(en/k) \leq m$, we have minimax risk estimate

$$\inf_{\widehat{\boldsymbol{u}}}\sup_{\boldsymbol{u}\in\mathcal{S}_k}\mathbb{E}\left[\|\widehat{\boldsymbol{u}}-\boldsymbol{u}\|_2^2\right] = \Theta\left(\frac{\sigma_z^2 k\log(en/k)}{m}\right).$$

For each $1 \le l \le L$, let $\widehat{\boldsymbol{u}}_l$ be the minimax estimator from Definition G.3, and let $\widehat{\boldsymbol{x}}^2 := \frac{1}{L}\sum_{l=1}^L \widehat{\boldsymbol{u}}_l^2$ be our estimator for the unknown signal $\boldsymbol{x}_o \in \mathcal{S}_k^{\mathrm{bdd}}$. It follows that

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\boldsymbol{x}}^2 - \boldsymbol{x}_o^2\right\|_2^2\right] =& \mathbb{E}\left[\left\|\widehat{\boldsymbol{x}}^2 - \frac{1}{L}\sum_{l=1}^L \boldsymbol{u}_l^2 + \frac{1}{L}\sum_{l=1}^L \boldsymbol{u}_l^2 - \boldsymbol{x}_o^2\right\|_2^2\right] \\
\overset{(a)}{\le}& 2\mathbb{E}\left[\left\|\widehat{\boldsymbol{x}}^2 - \frac{1}{L}\sum_{l=1}^L \boldsymbol{u}_l^2\right\|_2^2 + \left\|\frac{1}{L}\sum_{l=1}^L \boldsymbol{u}_l^2 - \boldsymbol{x}_o^2\right\|_2^2\right] \\
=& \frac{2}{L^2}\mathbb{E}\left[\left\|\sum_{l=1}^L \left(\widehat{\boldsymbol{u}}_l^2 - \boldsymbol{u}_l^2\right)\right\|_2^2\right] + \frac{2}{L^2}\mathbb{E}\left[\left\|\sum_{l=1}^L \left(\boldsymbol{u}_l^2 - \boldsymbol{x}_o^2\right)\right\|_2^2\right].
\end{aligned}$$

where for (a) we used $(a+b)^2 \le 2(a^2+b^2)$ in the vector form.

Since $(\boldsymbol{u}_l^2 - \boldsymbol{x}_o^2)$'s are k-sparse, independent and mean zero, for the second term we have

$$\frac{2}{L^2}\mathbb{E}\left[\left\|\sum_{l=1}^L \left(\boldsymbol{u}_l^2 - \boldsymbol{x}_o^2\right)\right\|_2^2\right] = \frac{2}{L^2}\sum_{l=1}^L \mathbb{E}\left[\left\|\boldsymbol{u}_l^2 - \boldsymbol{x}_o^2\right\|_2^2\right] \le \frac{8x_{\max}^2 \mathbb{E}\left[(\zeta^2-1)^2\right] k}{L}.$$

where $\zeta \sim N(0,1)$.

To treat the first term, we notice that by Definition G.3

$$\frac{2}{L^2}\mathbb{E}\left[\left\|\sum_{l=1}^L \left(\widehat{\boldsymbol{u}}_l^2 - \boldsymbol{u}_l^2\right)\right\|_2^2\right] \overset{(a)}{\le} \frac{2}{L^2}\cdot L\sum_{l=1}^L \mathbb{E}\left[\left\|\widehat{\boldsymbol{u}}_l^2 - \boldsymbol{u}_l^2\right\|_2^2\right] \le \frac{2\sigma_z^2 k \log(en/k)}{m}.$$

where for (a) we have used the elementary inequality $(a_1 + \cdots + a_L)^2 \le L(a_1^2 + \cdots + a_L^2)$.