Perceptual adjustment queries: An inverted measurement paradigm for low-rank metric learning

Austin Xu [1] **Andrew D. McRae** [2] **Jingyan Wang** [1] **Mark A. Davenport** [1] **Ashwin Pananjady** [1]

Abstract

We introduce a new type of informative and yet cognitively lightweight query mechanism for collecting human feedback, called the perceptual adjustment query (PAQ). The PAQ combines advantages from both ordinal and cardinal queries. We showcase the PAQ mechanism by collecting observations on a metric space involving an unknown Mahalanobis distance, and consider the problem of learning this metric from PAQ measurements. This gives rise to a type of high dimensional, low-rank matrix estimation problem under a new measurement scheme to which standard matrix estimators cannot be applied. Consequently, we develop a two-stage estimator for metric learning from PAQs, and provide sample complexity guarantees for this estimator. We demonstrate the performance along with various properties of the estimator by extensive numerical simulations.

1. Introduction

Should we query cardinal or ordinal data from people? This question arises in a broad range of applications, such as in conducting surveys (Rankin & Grube, 1980; Harzing et al., 2009; Yannakakis & Hallam, 2011), grading assignments (Shah et al., 2013; Raman & Joachims, 2014), and evaluating employees (Goffin & Olson, 2011), to name a few. *Cardinal* data refer to numerical scores. For example, teachers score writing assignments in the range of 0-100, and survey respondents express their agreement with a statement on a scale of 1 to 7. *Ordinal* data refer to relations between items, such as pairwise comparisons (choosing the better item in a pair) and rankings (ordering all or a subset of items). There is no free lunch, and both cardinal and ordinal queries have pros and cons.

[1]Georgia Institute of Technology [2]EPFL. Correspondence to: Austin Xu <axu77@gatech.edu>.

The Many Facets of Preference Learning Workshop at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

On the one hand, collecting ordinal data is typically more efficient in terms of worker time and cognitive load (Shah et al., 2016), and surprisingly often matches or exceeds the accuracy of cardinal data (Rankin & Grube, 1980; Shah et al., 2016). The information contained in ordinal queries, however, is fundamentally limited and lacks expressiveness. On the other hand, cardinal data are more expressive (Wang & Shah, 2019). For example, scoring two items 1 and 2 conveys a very different message from scoring them 9 and 10 or 1 and 10, although all yield the same pairwise comparison outcome. However, the expressiveness of cardinal data often comes at the cost of miscalibration: Prior work has shown that different people have different scales (Griffin & Brenner, 2008), and even one person's scale can drift over time (e.g., Harik et al., 2009; Myford & Wolfe, 2009). These inter-person and intra-person discrepancies make it challenging to interpret and aggregate raw scores effectively.

The goal of this paper is to study whether one can combine the advantages of cardinal and ordinal queries to achieve the best of both worlds. Specifically, we pose the research question:

> *Can we develop a new paradigm for human data elicitation that is expressive, accurate, and cognitively lightweight?*

Towards this goal, we extract key features of both cardinal and ordinal queries, and propose a new type of query scheme that we term the *perceptual adjustment query* (PAQ). As a thought experiment, consider the task of learning an individual's preferences. The query can take the following forms:

- **Ordinal:** Do you prefer a \$2 bus ride that takes 40 minutes or a \$25 taxi that takes 10 minutes?
- **Cardinal:** On a scale of 0 to 1, how much do you value a \$2 bus ride that takes 40 minutes?
- **Proposed approach:** To reach the same level of preference for a \$2 bus trip that takes 40 minutes, a taxi that takes 10 minutes would cost $\$x$.

A user interface for the proposed approach is shown in Figure 1 (top). We present the user a reference item (a \$2 bus

Figure 1: The user interface for preference learning (top) and similarity learning (bottom).

ride that takes 40 minutes), and a sliding bar representing the number of dollars (x) for the 10 minute taxi cost. As the user adjusts the slider, the value of x starts with 0 and gradually increases on a continuous scale. The user is instructed to place the slider at a point where they equally prefer a $2 bus ride and a taxi ride of x dollars. The PAQ thus combines cardinal and ordinal elicitation, by asking the user to make cognitive judgments in a relative sense by comparing items, while reporting a numerical value derived from the location of the slider. The ordinal reasoning endows the query with accuracy and efficiency, while the cardinal output enables a more expressive response. Moreover, this cardinal output mitigates miscalibration, because instead of asking to rate on a subjective and ambiguous notion (i.e., preference), we provide the user a reference object to anchor their rating scale (i.e., the taxi ride amount that gives the same preference as the bus ride).

In this paper, we apply the PAQ scheme in the framework of metric learning for human perception. In this model, items are represented by points in a (possibly high-dimensional) space, and the goal is to learn a distance metric such that a smaller distance between a pair of items means that they are semantically and perceptually closer, and vice versa. Figure 1 (bottom) presents a PAQ for collecting similarity data for metric learning, where the user is instructed to place the slider at the precise point where the object appears to transition from being similar to dissimilar. This sequence of images could be generated by traversing a path in the latent space of a generative model — given a latent feature vector, the model can synthesize a corresponding image.

This example showcases two additional advantages of the PAQ mechanism. First, PAQs provide "hard examples" by design and thus allow effective learning. Consider Figure 1 (bottom): Items on the left of the spectrum are apples (clearly similar to the reference), and items on the right are pears (clearly dissimilar to the reference), and only a small subset of items in the middle appear ambiguous. PAQs collect information precisely about "confusing" items in this ambiguous region. On the other hand, if a pairwise comparison samples the target item uniformly at random from the ones shown, it rarely falls in the ambiguous region. Without information about confusing items, it is difficult to learn a metric capable of disambiguating categories of items.

A second advantage is that the PAQ provides users with the *context* of a specific dimension along which items vary. For example, consider a pairwise comparison between the reference item and the "yellow apple " selected in Figure 1. They have similar shapes, but different colors, so the user lacks context to judge whether they should be considered similar or dissimilar. In contrast, the full spectrum provided in the PAQ tells the user that the similarity judgment is apples vs. pears. The access to such context improves self-consistency in user responses (Canal et al., 2020).

1.1. Our contributions

We propose the *perceptual adjustment query* (PAQ), which combines cardinal reporting and ordinal reasoning using a sliding bar for relational questions. We demonstrate the applicability of this query to metric learning under a Mahalanobis metric. We first present a mathematical formulation of this in Section 2. We then show that the sliding bar response can be viewed as an *inverted measurement* of the metric matrix that we want to estimate, and this allows us to restate our problem as that of estimating a low-rank matrix from a specific type of trace measurements (Section 3). However, our PAQ formulation differs from classical matrix estimation due to two technical challenges: (a) the sensing matrices and noise are correlated, and (b) the sensing matrices are heavy-tailed. We propose a query procedure and estimator that overcome these two challenges via *sample averaging* and *truncation* (Section 3), and we prove statistical error bounds on the estimator error (Section 4). The unconventional nature of the sensing model and estimator causes surprising behaviors in our error bounds; we present simulations verifying that these behaviors in practice in Section 5.

1.2. Related work

In metric learning, prior work considers using paired comparison (of the form "are these two items similar or dissimilar") (Ying et al., 2009; Bian & Tao, 2012; Guo & Ying, 2014; Bellet & Habrard, 2015) and triplet comparisons (of the form "which of the two items $\boldsymbol{x}$ and $\boldsymbol{x}'$ is more similar to the reference item $\boldsymbol{x}_0$?") (Mason et al., 2017). The metric learning from triplets problem is generalized by Xu & Davenport (2020) to consider an *unknown* reference point (referred to as an "ideal point") that captures different individual preferences. Tuple queries (Canal et al., 2020) extends triplets to ranking more than two items with respect to a reference item. PAQ can be viewed as extending this set of items to a continuous spectrum. However, the goal of tuple queries is to rank the items, whereas in PAQ the

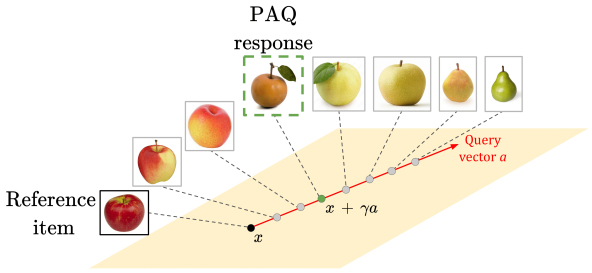


Figure 2: Given a reference item $\boldsymbol{x}$ and a query vector $\boldsymbol{a}$, a continuous path of items is formed $\{\boldsymbol{x} + \gamma \boldsymbol{a} : \gamma \in [0, \infty)\}$. A user is asked to pick the first item along this path that is dissimilar to the reference item, denoted $\boldsymbol{x} + \gamma \boldsymbol{a}$.

ranking is provided by the feature space and we ask people to identify a transition point (similar vs. dissimilar) in this ranking.

Our theoretical formulation (presented in Section 3) resembles the problem of low-rank matrix estimation from trace measurements (e.g., Recht et al., 2010; Negahban & Wainwright, 2011; Tsybakov & Rohde, 2011; Candes & Plan, 2011; Negahban et al., 2012; Cai & Zhang, 2013), and in particular, when the sensing matrix is of rank one (Cai & Zhang, 2015; Chen et al., 2015; Kueng et al., 2017; McRae et al., 2022). However, our model presents two important distinctions from prior literature. In our case, the sensing matrices are both heavy-tailed and correlated with the measurement noise. The heavy-tailed matrices violate the assumptions of much prior work that relies on sub-Gaussian or sub-exponential assumptions on the sensing matrices. We draw particular inspiration from Fan et al. (2021), which studies applying truncation to control heavy-tailed behavior in a number of problem settings. However, in the low-rank matrix estimation setting, Fan et al. (2021) only analyzes the case of heavy-tailed noise under a sub-Gaussian design, meaning their methodology and results are not applicable to our problem setting.

2. Model

In this section, we present our model for the perceptual adjustment query (PAQ) in metric learning.

2.1. Mahalanobis metric learning

We consider a d-dimensional feature space: thus each item is represented by a point in $\mathbb{R}^d$. The distance metric model for human similarity perception posits that there is a metric on $\mathbb{R}^d$ that measures how dissimilar items are perceived to be. A recent line of work (Xu & Davenport, 2020; Canal et al., 2022) has modeled the distance metric as a Mahalanobis metric. If $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is a symmetric positive semi-definite (PSD) matrix, the squared Mahalanobis distance with respect to $\boldsymbol{\Sigma}$ between items $\boldsymbol{x}$ and $\boldsymbol{x}' \in \mathbb{R}^d$ is $\|\boldsymbol{x} - \boldsymbol{x}'\|_{\boldsymbol{\Sigma}}^2 := (\boldsymbol{x} - \boldsymbol{x}')^\top \boldsymbol{\Sigma} (\boldsymbol{x} - \boldsymbol{x}')$. The distance represents the extent of dissimilarity between items $\boldsymbol{x}$ and $\boldsymbol{x}'$: if we further have a perceptual boundary value $y > 0$, this model posits that items $\boldsymbol{x}, \boldsymbol{x}'$ are perceived as similar if $\|\boldsymbol{x} - \boldsymbol{x}'\|_{\boldsymbol{\Sigma}}^2 < y$ and dissimilar if $\|\boldsymbol{x} - \boldsymbol{x}'\|_{\boldsymbol{\Sigma}}^2 > y$.

In Mahalanobis metric learning, we assume that perception is (approximately) governed by an unknown "ground truth" matrix $\boldsymbol{\Sigma}^\star$, and our goal is to estimate $\boldsymbol{\Sigma}^\star$ from user feedback. Note that the problem is scale-invariant, in the sense that for any pair $(\boldsymbol{\Sigma}, y)$, the similarity predictions are exactly the same as those made by the pair $(c\boldsymbol{\Sigma}, cy)$ for any constant $c > 0$. Hence, one can set y to be any positive value without loss of generality. We adopt a high-dimensional framework and, following existing work (Mason et al., 2017; Canal et al., 2022), assume that the matrix $\boldsymbol{\Sigma}^\star$ is low-rank.

2.2. The perceptual adjustment query (PAQ)

We assume that every point in our feature space $\mathbb{R}^d$ corresponds to some item (e.g., if the space is the latent space of a generative model). Recall from Figure 1 that a PAQ collects similarity data between a pair of items, where a reference item is fixed, and a spectrum of target items is generated from a one dimensional path in the feature space. Denote the reference item by $\boldsymbol{x} \in \mathbb{R}^d$. The target items can be generated by any path in $\mathbb{R}^d$, but for simplicity, we consider straight lines. For any vector $\boldsymbol{a} \in \mathbb{R}^d$, we construct the line $\{\boldsymbol{x} + \gamma \boldsymbol{a} : \gamma \in [0, \infty)\}$. We call this vector $\boldsymbol{a}$ the query vector. The PAQ traverses this line by increasing the value of γ starting from 0, as shown in Figure 2. In the user interface, γ represents the distance between the leftmost possible location of the slider, and its current location, i.e., the length that the slider has traversed.

The PAQ instructs the user to slide to the transition point where the target item transitions from being similar to dissimilar with the reference item. According to our model, this transition point occurs when the $\boldsymbol{\Sigma}^\star$-Mahalanobis distance between the target item and the reference item is y. According to our model, the *ideal* location, denoted by $\gamma_\star$, at which the user stops the slider satisfies the equation

$$y = \|\boldsymbol{x} - (\boldsymbol{x} + \gamma_\star \boldsymbol{a})\|_{\boldsymbol{\Sigma}^\star}^2 = \gamma_\star^2 \boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}. \tag{1}$$

Note that the ideal PAQ response $\gamma_\star$ does not depend on the specific reference item $\boldsymbol{x}$ but rather only on the query direction $\boldsymbol{a}$ and the (unknown) metric matrix $\boldsymbol{\Sigma}^\star$. When querying users with PAQs, the practitioner has control over how the query vectors $\boldsymbol{a}$ are selected. We discuss how to select $\boldsymbol{a}$ in Section 3.2.

2.3. Noise model

Human responses are noisy. We model this noise as follows: in the PAQ response equation (1), we replace the

boundary value y by $y + \eta$, where $\eta \in \mathbb{R}$ represents noise. Thus the user provides a noisy response γ whose value satisfies $\gamma^2 \boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a} = y + \eta$. This noise model implies that if $\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}$ is large, then in the user interface Figure 1 (bottom), the semantic meaning of the item changes rapidly as the user moves the slider along the direction $\boldsymbol{a}$. In this case, the slider ends up within a small interval around the true transition point. On the other hand, if $\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}$ is small, then the image changes slowly as the user moves the slider. In this case, it is hard to distinguish where exactly the transition occurs, so the slider ends up in a larger interval around the transition point.

3. Methodology

In this section, we formally present the statistical estimation problem for metric learning from PAQ data, and we develop our algorithm for estimating the true metric matrix $\boldsymbol{\Sigma}^\star$.

3.1. Statistical estimation

Assume we make N PAQ response, using N query vectors $\{\boldsymbol{a}_i\}_{i=1}^N$ that we can specify. Denote the noise associated with these queries by random variables $\eta_1, \ldots, \eta_N \in \mathbb{R}$. We obtain PAQ responses, denoted by $\gamma_1, \ldots, \gamma_N$, that satisfy

$$\gamma_i^2 \boldsymbol{a}_i^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}_i = y + \eta_i, \quad i = 1, \ldots, N. \tag{2}$$

We assume the noise variable η is independent.

Given the query directions $\{\boldsymbol{a}_i\}_{i=1}^N$ and the PAQ responses $\{\gamma_i\}_{i=1}^N$, we want to estimate the matrix $\boldsymbol{\Sigma}^\star$. We first rewrite our measurement model as follows: denote the matrix inner product by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle := \operatorname{tr}\left(\boldsymbol{A}^\top \boldsymbol{B}\right)$ for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of compatible dimension. Then, from (2), we write

$$\gamma^2 = \frac{y + \eta}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} = \frac{y + \eta}{\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{\Sigma}^\star \rangle}. \tag{3}$$

Plugging this once more into (2), we have

$$y + \eta = \langle \boldsymbol{A}^{\mathsf{inv}}, \boldsymbol{\Sigma}^\star \rangle,$$

where

$$\boldsymbol{A}^{\mathsf{inv}} := \gamma^2 \boldsymbol{a}\boldsymbol{a}^\top = \frac{y + \eta}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \boldsymbol{a}\boldsymbol{a}^\top. \tag{4}$$

Hence, our problem resembles trace regression, and, in particular, low-rank matrix estimation from rank-one measurements (because the matrix $\boldsymbol{A}^{\mathsf{inv}}$ has rank 1) (Cai & Zhang, 2015; Chen et al., 2015; Kueng et al., 2017; McRae et al., 2022). We call $\boldsymbol{A}^{\mathsf{inv}}$ the sensing matrix, and $\boldsymbol{a}$ the sensing vector. Classical trace regression assumes we make (noisy) observations of the form $y \approx \langle \boldsymbol{A}, \boldsymbol{\Sigma}^\star \rangle$ where $\boldsymbol{A}$ is fixed before we make the measurement; in our problem, the sensing matrix $\boldsymbol{A}^{\mathsf{inv}}$ depends on our observed response γ and associated sensing vector $\boldsymbol{a}$. Hence, the process of obtaining a PAQ response can be viewed as an *inversion* of the standard trace measurement process. The inverse nature of our problem makes estimator design more challenging, as we discuss in the following section.

3.2. Algorithm

As our first attempt at a procedure to estimate $\boldsymbol{\Sigma}^\star$, we follow the literature (Negahban & Wainwright, 2011; McRae et al., 2022) and consider randomly sampling i.i.d. vectors $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, I_d)$. We use standard least-squares estimation of $\boldsymbol{\Sigma}^\star$, with nuclear-norm regularization to promote low rank. We solve the following program:

$$\min_{\boldsymbol{\Sigma} \succeq \boldsymbol{0}} \frac{1}{N} \sum_{i=1}^N \left(y - \langle \boldsymbol{A}_i^{\mathsf{inv}}, \boldsymbol{\Sigma} \rangle\right)^2 + \lambda_N \|\boldsymbol{\Sigma}\|_*, \tag{5}$$

where $\|\cdot\|_*$ is the nuclear norm, and $\lambda_N > 0$ is a regularization parameter. This is a convex semidefinite program and can be solved with standard off-the-shelf solvers. However, the inverted form of our measurement model creates two critical issues when naïvely using (5):

- **Dependence between the sensing matrix and noise.** Note that the sensing matrix (4) depends on the noise η. Quantitatively, we have $\mathbb{E}\left[\eta \boldsymbol{A}^{\mathsf{inv}}\right] \neq 0$ (see Appendix A.1). Standard trace regression analyses require that this quantity be zero, typically assuming (at least) that η is zero-mean conditioned on the sensing matrix $\boldsymbol{A}$. The failure of this to hold in our case introduces a bias that does not decrease with the sample size N.
- **Heavy-tailed sensing matrix.** The factor $\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}$ in $\boldsymbol{A}^{\mathsf{inv}}$ makes $\boldsymbol{A}^{\mathsf{inv}}$ heavy-tailed in general. When $\boldsymbol{a}$ is Gaussian, the term $\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}$ is an inverse weighted chi-square random variable, which has infinite higher moments (the number of finite moments depends on the rank of $\boldsymbol{\Sigma}^\star$). This makes error analysis much more difficult, as standard analyses require the sensing matrix $\boldsymbol{A}$ to concentrate well.

To overcome these challenges, we make two key modifications to the procedure (5).

Step 1: Bias reduction via averaging. First, we want to mitigate the bias due to the dependence between the sensing matrix $\boldsymbol{A}^{\mathsf{inv}}$ and the noise η. The bias term $\mathbb{E}\left[\eta \boldsymbol{A}^{\mathsf{inv}}\right]$ scales proportionally to $\mathbb{E}\left[(y + \eta)\eta\right] = \mathbb{E}\left[\eta^2\right]$. Therefore, to reduce this bias in the least-squares estimator (5), we need to reduce the noise variance. We reduce the effective noise variance (and hence the bias) by *averaging* i.i.d. samples. Operationally, instead of obtaining N measurements from N distinct sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^N$, we draw sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^n$, and collect m measurements, denoted by $\{\gamma_i^{(j)}\}_{j=1}^m$, corresponding to each sensing vector $\boldsymbol{a}_i$. We

refer to n as the number of sensing vectors. To keep the total number of measurements constant, we set $n = \frac{N}{m}$, where the value of m is specified later. For each sensing vector $\boldsymbol{a}_i$, we compute the empirical mean of the m measurements:

$$\bar{\gamma}_i^2 := \frac{1}{m}\sum_{j=1}^{m}(\gamma_i^{(j)})^2 = \frac{1}{m}\sum_{j=1}^{m}\frac{y+\eta_i^{(j)}}{\boldsymbol{a}_i^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}_i} = \frac{y+\bar{\eta}_i}{\boldsymbol{a}_i^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}_i}, \quad (6)$$

where we define the average noise by $\bar{\eta}_i := \frac{1}{m}\sum_{j=1}^{m}\eta_i^{(j)}$. This averaging operation reduces the effective noise variance from $\mathrm{var}(\eta_i) = \nu_\eta^2$ to $\mathrm{var}(\bar{\eta}_i) = \frac{\nu_\eta^2}{m}$.

Step 2: Heavy tail mitigation via truncation. Next, we need to control the heavy-tailed behavior introduced by the $\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}$ term in the sensing matrix $\boldsymbol{A}^{\text{inv}}$. Note that the sample averaging procedure (6) does not mitigate this problem. We adopt the approach in Fan et al. (2021) and truncate the observations. Specifically, we truncate the averaged measurements $\bar{\gamma}_i^2$ to the value $\widetilde{\gamma}_i^2 := \bar{\gamma}_i^2 \wedge \tau$, where $\tau > 0$ is a truncation threshold that we will specify. We then construct the truncated sensing matrices

$$\widetilde{\boldsymbol{A}}_i = \widetilde{\gamma}_i^2 \boldsymbol{a}_i \boldsymbol{a}_i^\top = \left(\frac{y+\bar{\eta}_i}{\boldsymbol{a}_i^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}_i} \wedge \tau\right)\boldsymbol{a}_i \boldsymbol{a}_i^\top, i = 1, \ldots, n. \quad (7)$$

While truncation mitigates heavy-tailed behavior, it also introduces additional bias in our estimate. The truncation threshold τ therefore gives us another tradeoff, and in our analysis to follow, we carefully set the value of τ to balance the effects of heavy-tailedness and bias.

Final algorithm. After these two steps, we substitute the averaged and truncated matrices $\{\widetilde{\boldsymbol{A}}_i\}_{i=1}^n$ into the original least-squares problem (5). Specifically, we solve for

$$\widehat{\boldsymbol{\Sigma}} \in \operatorname*{arg\,min}_{\boldsymbol{\Sigma}\succeq \boldsymbol{0}} \frac{1}{n}\sum_{i=1}^{n}\left(y - \langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{\Sigma}\rangle\right)^2 + \lambda_n \|\boldsymbol{\Sigma}\|_*, \quad (8)$$

where, again, λ_n is a regularization parameter that we will specify. The full query and estimation procedure is presented in Algorithm 1.

Practical considerations. In the averaging step, we collect m measurements for each sensing vector $\boldsymbol{a}_i$. These measurements could be collected from m different users. Furthermore, recall from Section 2.2 that the measurements do not depend on the reference item $\boldsymbol{x}$. As a result, one may also collect multiple responses from the same user by presenting them the same query vector $\boldsymbol{a}_i$ but with different reference items $\boldsymbol{x}$. In addition, recall from Section 2.1 that the problem is scale-invariant. Practitioners are hence free to set the boundary y to be any positive value of their choice without loss of generality, and the noise variance ν_η^2 scales accordingly with y. The user interface does not depend on the value of y.

Algorithm 1 Metric learning from PAQ.

Input: number of total measurements N, averaging parameter m, truncation threshold τ, measurement value y, regularization parameter λ.

Compute the number of sensing vectors $n = \frac{N}{m}$
for $i = 1$ **to** n **do**
  Draw $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$
  Obtain m PAQ measurements $\gamma_i^{(1)}, \ldots, \gamma_i^{(m)}$ using $\boldsymbol{a}_i$ and y
end for
for $i = 1$ **to** n **do**
  Compute averaged response $\bar{\gamma}_i^2 = \frac{1}{m}\sum_{j=1}^{m}(\gamma_i^{(j)})^2$
  Compute truncated response $\widetilde{\gamma}_i^2 = \bar{\gamma}_i^2 \wedge \tau$
  Compute truncated sensing matrices $\widetilde{\boldsymbol{A}}_i = \widetilde{\gamma}_i^2 \boldsymbol{a}_i \boldsymbol{a}_i^\top$
end for
Compute $\widehat{\boldsymbol{\Sigma}} \in \operatorname*{arg\,min}_{\boldsymbol{\Sigma}\succeq \boldsymbol{0}} \frac{1}{n}\sum_{i=1}^{n}\left(y - \langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{\Sigma}\rangle\right)^2 + \lambda\|\boldsymbol{\Sigma}\|_*$

Output: matrix $\widehat{\boldsymbol{\Sigma}}$.

4. Theoretical results

We now present our main theoretical result, which is a finite-sample error bound for estimating a low-rank metric from inverted measurements with the nuclear norm-based estimator (8). Our error bound depends on the averaging parameter m and truncation threshold τ. Recall that the direction vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are sampled i.i.d. from $\mathcal{N}(\boldsymbol{0}, I_d)$. We have furthermore assumed that for any query direction $\boldsymbol{a}$, the noise variable η in the generic observation model $y + \eta = \langle \gamma^2 \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{\Sigma}^\star\rangle$ is independent of $\boldsymbol{a}$, has variance ν_η^2, and is bounded as $-y \le \eta \le \eta^\uparrow$ for some constant $\eta^\uparrow \ge 0$.

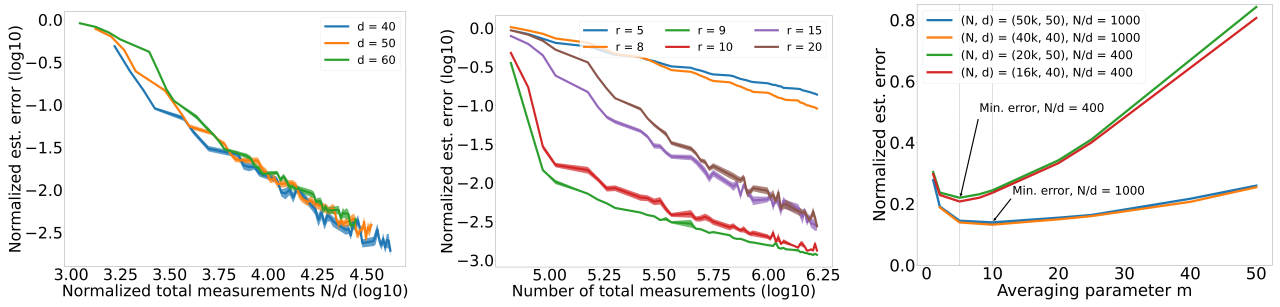
Theorem 1. *Let $y^\uparrow = y + \eta^\uparrow$ and κ_y be the median of $y + \bar{\eta}$. Suppose that $\boldsymbol{\Sigma}^\star$ has rank r, with $r > 8$. Denote by $\sigma_1 \ge \cdots \ge \sigma_r > 0$ the non-zero singular values of $\boldsymbol{\Sigma}^\star$. Assume that the truncation threshold τ satisfies $\tau \ge \frac{\kappa_y}{\mathrm{tr}(\boldsymbol{\Sigma}^\star)}$. Then there exist positive absolute constants c, C, C_1, and C_2 such that, if the regularization parameter and number of sensing vectors satisfy*

$$\lambda_n \ge C_1 y^\uparrow \left(\frac{y^\uparrow}{\sigma_r r}\sqrt{\frac{d}{n}} + \frac{d}{n}\tau + \left(\frac{y^\uparrow}{\sigma_r r}\right)^2 \frac{1}{\tau} + \frac{1}{\sigma_r r}\frac{\nu_\eta^2}{m}\right)$$
$$n \ge C_2 r d, \quad (9)$$

then any solution $\widehat{\boldsymbol{\Sigma}}$ to the optimization program (8) satisfies

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_F \le C\left(\frac{\mathrm{tr}\,(\boldsymbol{\Sigma}^\star)}{\kappa_y}\right)^2 \sqrt{r}\lambda_n \quad (10)$$

with probability at least $1 - 4\exp(-d) - \exp(-cn)$.



(a) (b) (c)

Figure 3: Simulations quantifying the effect of dimension d, rank r, and averaging parameter m on estimation error.

The proof of Theorem 1 is presented in Appendix B. The two sources of bias discussed in Section 3.2 appear in the expression (9) for the regularization parameter λ_n (and consequently in (10)). The term scaling as $1/\tau$ corresponds to the bias induced by truncation, and decreases as the truncation gets milder. The term scaling as ν_η^2/m corresponds to the bias from the noise–sensing matrix dependence. As discussed in Section 3.2, m-averaging results in a bias that scales like $1/m$. Given the dependence of the estimation error bound on the parameters m and τ, we carefully set these to obtain the tightest possible bound as a function of the number of *total measurements* $N = mn$.

Corollary 1. *Recall $N = mn$. Assume the conditions of Theorem 1 hold, that $N \gtrsim \nu_\eta^2 r^{3/2} d$, and that $(\nu_\eta^2)^2 N/d \geq 1$. Set the averaging parameter m and truncation threshold τ to be*

$$m = \left\lfloor \left(\frac{(\nu_\eta^2)^2 N}{d} \right)^{1/3} \right\rfloor \quad \text{and} \quad \tau = \frac{y^\uparrow}{\sigma_r r} \sqrt{\frac{N}{md}}. \tag{11}$$

Then setting λ_n equal to its lower bound in (9), any solution $\widehat{\boldsymbol{\Sigma}}$ to (8) satisfies

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_F \leq C \frac{\sigma_1^2}{\sigma_r} \left(\frac{y^\uparrow}{\kappa_y} \right)^2 r^{3/2} \left(\frac{\nu_\eta^2 d}{N} \right)^{1/3} \tag{12}$$

with probability at least $1 - 4\exp(-d) - \exp(-cN/m)$.

Corollary 1 follows immediately from Theorem 1. We provide a proof for completeness in Appendix E. Under the standard trace measurement model, if the measurement matrices are i.i.d. according to some sub-Gaussian distribution and $N \gtrsim rd$, then nuclear norm regularized estimators achieve an error that scales as $\sqrt{\frac{rd}{N}}$ (e.g., Negahban & Wainwright, 2011; Tsybakov & Rohde, 2011). Allowing heavier-tailed assumptions on the sensing matricestypically results in additional $\log d$ factors but does not impact the square-root rate (Negahban & Wainwright, 2012; Kueng et al., 2017; Fan et al., 2021). However, a crucial assumption in these results is that $\mathbb{E}[\eta \boldsymbol{A}] = \mathbf{0}$, and thus there is no bias due to measurement noise. Our inverted measurement sensing matrix is heavy-tailed and leads to measurement noise bias, which results in the one-third error rate.

5. Numerical simulations

In this section, we provide numerical simulations. For all results, we report the normalized estimation error $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_F / \|\boldsymbol{\Sigma}^\star\|_F$ averaged over 20 trials. Shaded areas (sometimes not visible) represent standard error of the mean. For all experiments, we follow Mason et al. (2017) and generate the ground truth matrix as $\boldsymbol{\Sigma}^\star = \frac{d}{\sqrt{r}} \boldsymbol{U}\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{d \times r}$ is a randomly generated matrix with orthonormal columns. The noise η is sampled from a uniform distribution on $[-\eta^\uparrow, \eta^\uparrow]$. We set the regularization parameter, truncation threshold, and averaging parameter consistent with our theoretical results (see (9) and (11)).

Effects of dimension and rank. We first characterize the effects of dimension d and matrix rank r. For all experiments, unless we are sweeping a specific parameter, we set $y = 200$, $d = 50$, $r = 15$, and $\eta^\uparrow = 10$. Fig. 3a shows the performance for varying values of d plotted against the normalized sample size N/d. For all dimensions d, the error decays to zero as the total number of measurements N increases. Furthermore, the error curves are well-aligned when the sample size is normalized by d; this agrees with the prediction of Corollary 1. Fig. 3b shows the performance for varying values of rank r. Recall that for our theoretical results we assume $r > 8$. When $r > 8$, the number of measurements required for the same estimation error increases as the rank increases. The error still decreases with N for $r \leq 8$, but at a markedly slower rate than when $r > 8$.

Effect of averaging parameter m. Equation (11) suggests that the averaging parameter m should scale proportionally to $(N/d)^{1/3}$. To test this, we set $y = 200$, $d = 50$, $r = 9$, and $\eta^\uparrow = 200$. We vary values of m for different choices of the (N, d) pair, as shown in Fig. 3c. The empirically optimal choice of m is observed to be the same when N/d is fixed, regardless of the particular choices of N or d.

References

Bao, Y. and Kan, R. On the moments of ratios of quadratic forms in normal random variables. *Journal of Multivariate Analysis*, 117:229–245, 2013.

Bellet, A. and Habrard, A. Robustness and generalization for metric learning. *Neurocomputing*, 151:259–267, 2015.

Bian, W. and Tao, D. Constrained empirical risk minimization framework for distance metric learning. *IEEE transactions on neural networks and learning systems*, 23 (8):1194–1205, 2012.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Cai, T. T. and Zhang, A. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE transactions on information theory*, 60(1):122–132, 2013.

Cai, T. T. and Zhang, A. Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

Canal, G., Fenu, S., and Rozell, C. Active ordinal querying for tuplewise similarity learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3332–3340. AAAI Press, 2020.

Canal, G., Mason, B., Vinayak, R. K., and Nowak, R. D. One for all: Simultaneous metric and preference learning over multiple users. In *Advances in Neural Information Processing Systems*, 2022.

Candes, E. J. and Plan, Y. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

Chen, Y., Chi, Y., and Goldsmith, A. J. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

Fan, J., Wang, W., and Zhu, Z. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, 49(3):1239, 2021.

Goffin, R. D. and Olson, J. M. Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, 6(1):48–60, 2011.

Griffin, D. and Brenner, L. *Perspectives on Probability Judgment Calibration*, chapter 9. Wiley-Blackwell, 2008. ISBN 9780470752937.

Guo, Z.-C. and Ying, Y. Guaranteed classification via regularized similarity learning. *Neural computation*, 26(3): 497–522, 2014.

Harik, P., Clauser, B., Grabovsky, I., Nungester, R., Swanson, D., and Nandakumar, R. An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, 46:43–58, 2009.

Harzing, A.-W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., Espejo, A., Ferreira, R., Giroud, A., Koester, K., Liang, Y.-K., Mockaitis, A., Morley, M. J., Myloni, B., Odusanya, J. O., O'Sullivan, S. L., Palaniappan, A. K., Prochno, P., Choudhury, S. R., Saka-Helmhout, A., Siengthai, S., Viswat, L., Soydas, A. U., and Zander, L. Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review*, 18(4):417–432, 2009. ISSN 0969-5931.

Kueng, R., Rauhut, H., and Terstiege, U. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.

Mason, B., Jain, L., and Nowak, R. Learning low-dimensional metrics. *Advances in neural information processing systems*, 30, 2017.

McRae, A. D., Romberg, J., and Davenport, M. A. Optimal convex lifted sparse phase retrieval and pca with an atomic matrix norm regularizer. *IEEE Transactions on Information Theory*, 2022.

Mendelson, S. Learning without concentration. *Journal of the ACM (JACM)*, 62(3):1–25, 2015.

Myford, C. M. and Wolfe, E. W. Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46(4):371–389, 2009.

Negahban, S. and Wainwright, M. J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. 2012.

Raman, K. and Joachims, T. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 1037–1046, New York, NY,

USA, 2014. Association for Computing Machinery. ISBN 9781450329569.

Rankin, W. L. and Grube, J. W. A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology*, 10(3):233–246, 1980.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Rudelson, M. and Vershynin, R. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., and Ramchandran, K. A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education*, 2013.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramch, K., ran, and Wainwright, M. J. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

Tropp, J. A. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pp. 67–101. Springer, 2015.

Tsybakov, A. and Rohde, A. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39(2):887–930, 2011.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

Wang, J. and Shah, N. B. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pp. 864–872, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

Xu, A. and Davenport, M. Simultaneous preference and metric learning from paired comparisons. *Advances in Neural Information Processing Systems*, 33, 2020.

Yannakakis, G. N. and Hallam, J. Ranking vs. preference: A comparative study of self-reporting. In D'Mello, S., Graesser, A., Schuller, B., and Martin, J.-C. (eds.), *Affective Computing and Intelligent Interaction*, pp. 437–446, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-24600-5.

Ying, Y., Huang, K., and Campbell, C. Sparse metric learning via smooth optimization. *Advances in neural information processing systems*, 22, 2009.

A. Preliminaries and Notation

In this section, we provide an overview of the key tools that are utilized in our proofs. We first introduce notation which is used throughout our proofs.

Notation. For two real numbers a and b, let $a \wedge b = \min\{a, b\}$. Given a vector $\boldsymbol{x} \in \mathbb{R}^d$, denote $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|_2$ as the ℓ_1 and ℓ_2 norm, respectively. Denote $\mathcal{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ to be the set of vectors with unit ℓ_2 norm. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, denote $\|\boldsymbol{A}\|_F$, $\|\boldsymbol{A}\|_*$, and $\|\boldsymbol{A}\|_{\text{op}}$ as the Frobenius norm, nuclear norm, and operator norm, respectively. Denote $\mathbb{S}^{d\times d} = \{\boldsymbol{A} \in \mathbb{R}^{d\times d} : \boldsymbol{A} = \boldsymbol{A}^\top\}$ to be the set of symmetric $d \times d$ matrices. Denote $\boldsymbol{A} \succeq \boldsymbol{0}$ to mean $\boldsymbol{A}$ is symmetric positive semi-definite. For $\boldsymbol{A} \succeq \boldsymbol{0}$, define the (pseudo-) inner product $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\boldsymbol{A}} = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y}$ and the associated (pseudo-) norm $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$. For matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d_1\times d_2}$, denote $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \text{tr}\left(\boldsymbol{A}^\top \boldsymbol{B}\right)$ as the Frobenius inner product.

We use the notation $f(x) \lesssim g(x)$ to denote that there exists some universal positive constant $c > 0$, such that $f(x) \leq c \cdot g(x)$, and use the notation $f(x) \gtrsim g(x)$ when $g(x) \lesssim f(x)$.

We define random matrices

$$\bar{\boldsymbol{A}} = \bar{\gamma}^2 \boldsymbol{a}\boldsymbol{a}^\top = \frac{y + \bar{\eta}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \boldsymbol{a}\boldsymbol{a}^\top \tag{13}$$

and

$$\widetilde{\boldsymbol{A}} = \widetilde{\gamma}^2 \boldsymbol{a}\boldsymbol{a}^\top = \left(\frac{y + \bar{\eta}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \wedge \tau\right) \boldsymbol{a}\boldsymbol{a}^\top \tag{14}$$

as the sensing matrix formed with the m-averaged responses $\bar{\gamma}$ and truncated responses $\widetilde{\gamma}$, respectively.

A.1. Inverted measurement sensing matrices result in estimation bias.

Recall from Equation (4) that the random sensing matrix $\boldsymbol{A}^{\text{inv}}$ takes the form

$$\boldsymbol{A}^{\text{inv}} = \frac{y + \eta}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \boldsymbol{a}\boldsymbol{a}^\top. \tag{15}$$

Standard trace regression analysis requires that the bias term $\mathbb{E}[\eta \boldsymbol{A}] = \boldsymbol{0}$, typically by assuming (at least) that η is zero-mean conditioned on the sensing matrix $\boldsymbol{A}$. The following lemma shows that the bias term associated with the inverted measurements sensing matrix $\boldsymbol{A}^{\text{inv}}$ is nonzero, resulting in biased estimation

Lemma 1. *Let $\boldsymbol{A}^{\text{inv}}$ be the random matrix defined in Eq. (4) and η be the measurement noise. Then,*

$$\mathbb{E}\left[\eta \boldsymbol{A}^{\text{inv}}\right] \neq \boldsymbol{0}. \tag{16}$$

The proof of Lemma 1 is provided in Appendix A.6.1. As a result, utilizing established low-rank matrix estimators will result in biased estimation.

A.2. Sub-exponential random variables.

Our analysis will depend on sub-exponential random variables, a class of random variables with heavier tails than Gaussian. While many definitions of sub-exponential random variables exist (see, for example, Vershynin, 2018, Chapter 2.7), we will make use of one particular property, presented below.

If X is a sub-exponential random variable, then there exists some constant c (only dependent on the distribution underlying the random variable X) such that for all integers $p \geq 1$,

$$\left(\mathbb{E}|X|^p\right)^{1/p} \leq cp. \tag{17}$$

A.3. Bernstein's inequality.

A key ingredient in our proofs is the well-known Bernstein's inequality, which is a concentration inequality for sums of independent sub-exponential random variables.

Lemma 2 (Bernstein's inequality, adapted from Boucheron et al., 2013, Theorem 2.10)**.** *Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume exist positive numbers u_1 and u_2 such that*

$$\mathbb{E}\left[X_i^2\right] \leq u_1 \quad \text{and} \quad \mathbb{E}\left[|X_i|^p\right] \leq \frac{p!}{2} u_1 u_2^{p-2} \text{ for all integers } p \geq 2, \tag{18a}$$

Then for all $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathbb{E}\left[X_i\right]\right)\right| \geq \sqrt{\frac{2u_1 t}{n}} + \frac{u_2 t}{n}\right) \leq 2\exp(-t). \tag{18b}$$

A.4. Moments of the ratios of quadratic forms.

Because the quadratic term $\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}$ appears in the denominator of our sensing matrices, our analysis depends on quantifying the moments of the ratios of quadratic forms. This is done in the following lemma.

Lemma 3. *There exists an absolute constant $c > 0$ such that the following is true. Let $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, $\boldsymbol{\Sigma}^\star \in \mathbb{R}^{d\times d}$ be any PSD matrix with rank r, and $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ be an arbitrary symmetric matrix.*

(a) Suppose that $r > 8$. Then we have

$$\mathbb{E}\left(\frac{1}{\boldsymbol{a}^T \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right)^4 \leq \frac{c}{\sigma_r^4 r^4}.$$

(b) Suppose that $r > 2$. Then we have

$$\mathbb{E}\left(\frac{\boldsymbol{a}^\top \boldsymbol{U} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right) \leq \frac{c}{\sigma_r r}\|\boldsymbol{U}\|_*. \tag{19}$$

The proof of Lemma 3 is presented in Appendix A.6.2.

A.5. A fourth moment bound for $\bar{\gamma}^2$.

Throughout our analysis, we will utilize the fact that $\bar{\gamma}^2$ has a bounded fourth-moment. This bound is characterized in the following lemma.

Lemma 4. *Assume $r > 8$. Then the bound*

$$\mathbb{E}\left[\left(\bar{\gamma}^2\right)^4\right] \lesssim \left(\frac{y + \eta^\uparrow}{\sigma_r r}\right)^4 \tag{20}$$

holds, where σ_r is the smallest non-zero singular value of $\boldsymbol{\Sigma}^\star$.

The proof of Lemma 4 is presented in Appendix A.6.3. For notational simplicity of the proofs, we denote $M = c\left(\frac{y+\eta^\uparrow}{\sigma_r r}\right)^4$.

A.6. Proofs of preliminary lemmas

In this section, we present proofs for preliminary lemmas from Appendices A.1, A.4, and A.5.

A.6.1. Proof of Lemma 1

We show that $\mathbb{E}\left[\eta \boldsymbol{A}^{\mathsf{inv}}\right] \neq \boldsymbol{0}$. Using the independence of η and $\boldsymbol{a}$ and the assumption that η is zero mean, we have

$$\mathbb{E}\left[\eta \boldsymbol{A}^{\mathsf{inv}}\right] = \mathbb{E}\left[\frac{\eta(y+\eta)}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \boldsymbol{a}\boldsymbol{a}^\top\right] \tag{21}$$

$$= \mathbb{E}\left[\eta(y+\eta)\right] \mathbb{E}\left[\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \boldsymbol{a}\boldsymbol{a}^\top\right] \tag{22}$$

$$= \nu_\eta^2 \, \mathbb{E}\left[\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \boldsymbol{a}\boldsymbol{a}^\top\right]. \tag{23}$$

This expectation is non-zero, as the random matrix $\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \boldsymbol{a}\boldsymbol{a}^\top$ is symmetric-positive semidefinite. Therefore, we have $\mathbb{E}\left[\eta \boldsymbol{A}^{\mathsf{inv}}\right] \neq \boldsymbol{0}$, as desired.

A.6.2. PROOF OF LEMMA 3

Without loss of generality, we assume that $\boldsymbol{\Sigma}^\star$ is diagonal for the remainder of this proof. To prove each part of Lemma 3, we utilize results on the moments of ratios of quadratic forms. For non-negative integers p and q, we first verify that the mixed moment $\mathbb{E}\left[\frac{(\boldsymbol{a}^\top \boldsymbol{U}\boldsymbol{a})^p}{(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a})^q}\right]$ exists. By Bao & Kan (2013, Proposition 1), the mixed moment exists if $\frac{r}{2} > q$. This is assumed to be true for all parts of Lemma 3.

By Bao & Kan (2013, Proposition 2), we have the following expression for the mixed moment $\mathbb{E}\left[\frac{(\boldsymbol{a}^\top \boldsymbol{U}\boldsymbol{a})^p}{(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a})^q}\right]$:

$$\mathbb{E}\left[\frac{(\boldsymbol{a}^\top \boldsymbol{U}\boldsymbol{a})^p}{(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a})^q}\right] = \frac{1}{\Gamma(q)} \int_0^\infty t^{q-1} \left|\boldsymbol{\Delta}_t\right| \mathbb{E}\left[\left(\boldsymbol{a}^\top \boldsymbol{\Delta}_t \boldsymbol{U} \boldsymbol{\Delta}_t \boldsymbol{a}\right)^p\right] dt, \tag{24}$$

where $\boldsymbol{\Delta}_t = (\boldsymbol{I}_d + 2t\boldsymbol{\Sigma}^\star)^{-1/2}$ and $|\boldsymbol{\Delta}_t|$ is the determinant of $\boldsymbol{\Delta}_t$. Our results will depend on characterizing $|\boldsymbol{\Delta}_t|$. We begin by noting that $\boldsymbol{\Delta}_t$ is a diagonal matrix of the following form

$$\boldsymbol{\Delta}_t = \begin{bmatrix} \frac{1}{(1+2t\sigma_1)^{1/2}} & & & & & \\ & \ddots & & & & \\ & & \frac{1}{(1+2t\sigma_r)^{1/2}} & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix}. \tag{25}$$

It follows that the determinant $|\boldsymbol{\Delta}_t|$ can be written as the product $|\boldsymbol{\Delta}_t| = \prod_{j=1}^r \frac{1}{(1+2t\sigma_r)^{1/2}}$. Furthermore, this product can be bounded as follows:

$$\frac{1}{(1+2t\sigma_1)^{r/2}} \leq |\boldsymbol{\Delta}_t| \leq \frac{1}{(1+2t\sigma_j)^{r/2}}. \tag{26}$$

We are now ready to prove parts (a) and (b).

Part (a). This case corresponds to the case where $p = 0$ and $q = 4$. Using the integral expression (24) and upper bound on determinant (26), with these values of p and q, we have

$$\mathbb{E}\left[\left(\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right)^4\right] = \frac{1}{\Gamma(4)} \int_0^\infty t^3 \left|\boldsymbol{\Delta}_t\right| dt \tag{27}$$

$$\leq \frac{1}{\Gamma(4)} \int_0^\infty t^3 \frac{1}{(1+2t\sigma_r)^{r/2}} dt \tag{28}$$

Making the substitution $s = 1 + 2t\sigma_r$, we can evaluate the integral as follows.

$$\mathbb{E}\left[\left(\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right)^4\right] \leq \frac{1}{2\Gamma(4)\sigma_r} \int_1^\infty \left(\frac{s-1}{2\sigma_r}\right)^3 \frac{1}{s^{r/2}} ds \tag{29}$$

$$\lesssim \frac{1}{\sigma_r^4} \int_1^\infty \frac{(s-1)^3}{s^{r/2}} ds \tag{30}$$

$$= \frac{1}{\sigma_r^4} \int_1^\infty \left(\frac{s^3}{s^{r/2}} - 3\frac{s^2}{s^{r/2}} + 3\frac{s}{s^{r/2}} - \frac{1}{s^{r/2}}\right) ds \tag{31}$$

$$= \frac{1}{\sigma_r^4} \left(\frac{2}{r-8} - \frac{6}{r-6} + \frac{6}{r-4} - \frac{2}{r-2}\right). \tag{32}$$

Therefore, we have that there exists some absolute constant c such that

$$\mathbb{E}\left[\left(\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right)^4\right] \leq \frac{c}{\sigma_r^4 r^4}, \tag{33}$$

as desired.

Part (b). This case corresponds to the case where $p = q = 1$. We begin again with the integral expression (24) and upper bound on determinant (26):

$$\mathbb{E}\left[\left(\frac{\boldsymbol{a}^\top \boldsymbol{U}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right)\right] = \frac{1}{\Gamma(1)} \int_0^\infty |\boldsymbol{\Delta}_t| \, \mathbb{E}\left[\boldsymbol{a}^\top \boldsymbol{\Delta}_t \boldsymbol{U} \boldsymbol{\Delta}_t \boldsymbol{a}\right] dt \tag{34}$$

$$\leq \frac{1}{\Gamma(1)} \int_0^\infty \frac{1}{(1+2t\sigma_r)^{r/2}} \mathbb{E}\left[\boldsymbol{a}^\top \boldsymbol{\Delta}_t \boldsymbol{U} \boldsymbol{\Delta}_t \boldsymbol{a}\right] dt \tag{35}$$

We now bound the expectation term $\mathbb{E}\left[\boldsymbol{a}^\top \boldsymbol{\Delta}_t \boldsymbol{U} \boldsymbol{\Delta}_t \boldsymbol{a}\right]$. Note that for $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, the expectation $\mathbb{E}\left[\boldsymbol{a}^\top \boldsymbol{B} \boldsymbol{a}\right] = \operatorname{tr}(\boldsymbol{B})$ for any symmetric matrix $\boldsymbol{B}$. Therefore, we have

$$\mathbb{E}\left[\boldsymbol{a}^\top \boldsymbol{\Delta}_t \boldsymbol{U} \boldsymbol{\Delta}_t \boldsymbol{a}\right] = \operatorname{tr}(\boldsymbol{\Delta}_t \boldsymbol{U} \boldsymbol{\Delta}_t) \tag{36}$$

$$\leq \|\boldsymbol{\Delta}_t \boldsymbol{U} \boldsymbol{\Delta}_t\|_*. \tag{37}$$

Above, we have used the fact that $\operatorname{tr}(\boldsymbol{B}) \leq \|\boldsymbol{B}\|_*$ for any symmetric matrix $\boldsymbol{B}$. By Hölder's inequality for Schatten-p norms, we have that $\|\boldsymbol{\Delta}_t \boldsymbol{U} \boldsymbol{\Delta}_t\|_* \leq \|\boldsymbol{\Delta}_t\|_{\mathrm{op}}^2 \|\boldsymbol{U}\|_*$. Because $\boldsymbol{\Delta}_t$ is diagonal and the entries of $\boldsymbol{\Delta}_t$ are bounded between 0 and 1, we can bound the operator norm as $\|\boldsymbol{\Delta}_t\|_{\mathrm{op}} \leq 1$. Therefore

$$\mathbb{E}\left[\boldsymbol{a}^\top \boldsymbol{\Delta}_t \boldsymbol{U} \boldsymbol{\Delta}_t \boldsymbol{a}\right] \leq \|\boldsymbol{U}\|_* \tag{38}$$

Substituting this upper bound for the expectation term into the integral, we obtain

$$\mathbb{E}\left[\frac{\boldsymbol{a}^\top \boldsymbol{U}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right] \leq \|\boldsymbol{U}\|_* \int_0^\infty \frac{1}{(1+2t\sigma_j)^{r/2}} dt. \tag{39}$$

Evaluating this integral, we have for some absolute constant c,

$$\mathbb{E}\left[\frac{\boldsymbol{a}^\top \boldsymbol{U}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right] \leq \frac{c}{\sigma_r r} \|\boldsymbol{U}\|_*, \tag{40}$$

as desired.

A.6.3. Proof of Lemma 4

By the bounded noise assumption, $y + \bar{\eta} \leq y + \eta^\uparrow$. Therefore, we have

$$\mathbb{E}\left[\left(\bar{\gamma}^2\right)^4\right] = \mathbb{E}\left[\left(\frac{y+\bar{\eta}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right)^4\right] \tag{41}$$

$$\leq (y+\bar{\eta})^4 \cdot \mathbb{E}\left[\left(\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right)^4\right]. \tag{42}$$

It therefore suffices to bound the fourth moment of $\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}$, which is done in Lemma 3. Therefore,

$$\mathbb{E}\left[\left(\frac{1}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right)^4\right] \lesssim \left(\frac{1}{\sigma_r r}\right)^4, \tag{43}$$

as desired.

B. Proof of Theorem 1

Our goal is to derive finite sample error bounds for the estimator in Equation (8). For our estimator, if the regularization parameter is set to be sufficiently large (which we will characterize later), then the error matrix is guaranteed to be in some *error set* $\mathcal{E}$. For rank r symmetric positive semidefinite matrices, the error set $\mathcal{E}$ can be characterized as (Negahban & Wainwright, 2011)

$$\mathcal{E} = \left\{ \boldsymbol{U} \in \mathbb{S}^{d\times d} : \|\boldsymbol{U}\|_* \leq 4\sqrt{2r}\|\boldsymbol{U}\|_F \right\}, \tag{44}$$

where recall that $\mathbb{S}^{d\times d}$ denotes the set of symmetric $d \times d$ matrices.

A key condition for estimation under these settings is to ensure that the shrunken sensing matrices satisfy a restricted strong convexity (RSC) condition over the error set $\mathcal{E}$. That is, we must show that there exists some positive constant κ such that

$$\frac{1}{n}\sum_{i=1}^{n}\langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{U}\rangle^2 \geq \kappa\|\boldsymbol{U}\|_F^2 \qquad \text{for all } \boldsymbol{U} \in \mathcal{E}. \tag{45}$$

We begin by stating a proposition that characterizes the deterministic upper bound on the estimation error.

Proposition 1 (Fan et al., 2021, Theorem 1 with $q = 0$)**.** *Suppose that $\boldsymbol{\Sigma}^\star$ has rank r and the shrunken sensing matrices satisfy the restricted strong convexity condition with positive constant κ. Then if the regularization parameter satisfies*

$$\lambda_n \geq 2\left\| \frac{1}{n}\sum_{i=1}^{n} y\widetilde{\boldsymbol{A}}_i - \frac{1}{n}\sum_{i=1}^{n}\langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{\Sigma}^\star\rangle \widetilde{\boldsymbol{A}}_i \right\|_{\text{op}} \tag{46a}$$

any optimal solution $\widehat{\boldsymbol{\Sigma}}$ of the optimization program (8) satisfies

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_F \leq \frac{32\sqrt{r}\lambda_n}{\kappa} \tag{46b}$$

This theorem is a special case of Theorem 1 in (Fan et al., 2021), which is in turn adapted from Theorem 1 in (Negahban & Wainwright, 2011) (see (Negahban & Wainwright, 2011) or (Fan et al., 2021) for the proof). Proposition 1 is a deterministic and nonasymptotic result and provides a roadmap for proving upper bounds. First, we show that the operator norm (46a) can be upper bounded with high probability, allowing us to set the regularization parameter λ_n accordingly. Second, we show that the RSC condition (45) is satisfied with high probability. We begin by bounding the operator norm (46a) in the following proposition.

Proposition 2. *Let $y^\uparrow = y + \eta^\uparrow$. Suppose that $\boldsymbol{\Sigma}^\star$ has rank r, with $r > 8$. Then there exists a positive absolute constant C such that the bound*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} y\widetilde{\boldsymbol{A}}_i - \frac{1}{n}\sum_{i=1}^{n}\langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{\Sigma}^\star\rangle \widetilde{\boldsymbol{A}}_i \right\|_{\text{op}} \leq Cy^\uparrow\left(\frac{y^\uparrow}{\sigma_r r}\sqrt{\frac{d}{n}} + \frac{d}{n}\tau + \left(\frac{y^\uparrow}{\sigma_r r}\right)^2 \frac{1}{\tau} + \frac{1}{\sigma_r r}\frac{\nu_\eta^2}{m} \right) \tag{47}$$

holds with probability at least $1 - 4\exp(-d)$.

The proof of Proposition 2 is provided in Appendix B.1. Next, we show that the RSC condition (45) is satisfied with high probability, as is done in the following proposition.

Proposition 3. *Let κ_y be the median of $y + \bar{\eta}$ and let $\mathcal{E}$ be the error set defined in Eq. (44). Suppose that the truncation threshold τ satisfies $\tau \geq \frac{\kappa_y}{\text{tr}(\boldsymbol{\Sigma}^\star)}$. Then, there exist positive absolute constants $\kappa_{\mathcal{L}}$, c, and C such that if the number of effective measurements satisfy*

$$n \geq Crd \tag{48a}$$

then we have

$$\frac{1}{n}\sum_{i=1}^{n}\langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{U}\rangle^2 \geq \kappa_{\mathcal{L}}\left(\frac{\kappa_y}{\text{tr}\left(\boldsymbol{\Sigma}^\star\right)}\right)^2 \|\boldsymbol{U}\|_F^2 \tag{48b}$$

simultaneously for all matrices $\boldsymbol{U} \in \mathcal{E}$ with probability greater than $1 - \exp(-cn)$.

The proof of Proposition 3 is provided in Appendix B.2. We now utilize the results of Propositions 1, 2 and 3 to derive our final error bounds. By Proposition 2, we know that the operator norm (46a) can be upper bounded with high probability. We set the regularization parameter λ_n to satisfy

$$\lambda_n \geq C_1 y^{\uparrow} \left(\frac{y^{\uparrow}}{\sigma_r r} \sqrt{\frac{d}{n}} + \frac{d}{n}\tau + \left(\frac{y^{\uparrow}}{\sigma_r r} \right)^2 \frac{1}{\tau} + \frac{1}{\sigma_r r} \frac{\nu_\eta^2}{m} \right). \tag{49}$$

for an appropriate constant C_1. Furthermore, by Proposition 3, we have that there exists some universal constant C_2 such that if the number of effective measurements satisfies $n \geq C_2 r d$, the RSC condition also holds for constant $\kappa = \kappa_{\mathcal{L}} \left(\frac{\kappa_y}{\mathrm{tr}(\mathbf{\Sigma}^\star)} \right)^2$ with high probability. Taking a union bound, we have that Proposition 2 and Proposition 3 hold simultaneously with probability at least $1 - 4\exp(-d) - \exp(-cn)$. By Proposition 1, the bound

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star\|_F \leq 32\sqrt{r} \frac{\lambda_n}{\kappa_{\mathcal{L}} \left(\frac{\kappa_y}{\mathrm{tr}(\mathbf{\Sigma}^\star)} \right)^2} \tag{50}$$

$$\leq C \left(\frac{\mathrm{tr}\left(\mathbf{\Sigma}^\star\right)}{\kappa_y} \right)^2 \sqrt{r} \lambda_n \tag{51}$$

holds with probability at least $1 - 4\exp(-d) - \exp(-cn)$, as desired. Above, we have defined $C = \frac{32}{\kappa_{\mathcal{L}}}$.

B.1. Proof of Proposition 2.

Our goal is to derive an upper bound on the operator norm

$$\left\| \frac{1}{n} \sum_{i=1}^n y \widetilde{\boldsymbol{A}}_i - \frac{1}{n} \sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \mathbf{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}}_i \right\|_{\mathrm{op}}. \tag{52}$$

Step 1: decompose the error into five terms. We begin by adding and subtracting multiple quantities, as done below.

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n y \widetilde{\boldsymbol{A}}_i - \frac{1}{n} \sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \mathbf{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}}_i &= \frac{1}{n} \sum_{i=1}^n y \widetilde{\boldsymbol{A}}_i - \mathbb{E}\left[y \widetilde{\boldsymbol{A}} \right] + \mathbb{E}\left[y \widetilde{\boldsymbol{A}} \right] - \mathbb{E}\left[y \bar{\boldsymbol{A}} \right] \\ &\quad + \mathbb{E}\left[y \bar{\boldsymbol{A}} \right] - \mathbb{E}\left[\langle \widetilde{\boldsymbol{A}}, \mathbf{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}} \right] + \mathbb{E}\left[\langle \widetilde{\boldsymbol{A}}, \mathbf{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}} \right] - \frac{1}{n} \sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \mathbf{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}}_i \qquad (53) \\ &\overset{(i)}{=} \frac{1}{n} \sum_{i=1}^n y \widetilde{\boldsymbol{A}}_i - \mathbb{E}\left[y \widetilde{\boldsymbol{A}} \right] + \mathbb{E}\left[y \widetilde{\boldsymbol{A}} \right] - \mathbb{E}\left[y \bar{\boldsymbol{A}} \right] \\ &\quad + \mathbb{E}\left[\langle \bar{\boldsymbol{A}}, \mathbf{\Sigma}^\star \rangle \bar{\boldsymbol{A}} \right] - \mathbb{E}\left[\langle \widetilde{\boldsymbol{A}}, \mathbf{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}} \right] - \mathbb{E}\left[\bar{\eta} \bar{\boldsymbol{A}} \right] \\ &\quad + \mathbb{E}\left[\langle \widetilde{\boldsymbol{A}}, \mathbf{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}} \right] - \frac{1}{n} \sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \mathbf{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}}_i. \qquad (54) \end{aligned}$$

Above, (i) follows from substituting in $\langle \bar{\boldsymbol{A}}, \boldsymbol{\Sigma}^\star \rangle - \bar{\eta}$ for y for the $\mathbb{E}\left[y\bar{\boldsymbol{A}}\right]$ term. To obtain our final bound, we bound the following operator norms.

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^n y\widetilde{\boldsymbol{A}}_i - \frac{1}{n}\sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}}_i \right\|_{\text{op}} \leq & \underbrace{y\left\| \frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{A}}_i - \mathbb{E}\left[\widetilde{\boldsymbol{A}}\right] \right\|_{\text{op}}}_{\text{Term 1}} \\ & + \underbrace{y\left\| \mathbb{E}\left[\widetilde{\boldsymbol{A}}\right] - \mathbb{E}\left[\bar{\boldsymbol{A}}\right] \right\|_{\text{op}}}_{\text{Term 2}} \\ & + \underbrace{\left\| \mathbb{E}\left[\langle \bar{\boldsymbol{A}}, \boldsymbol{\Sigma}^\star \rangle \bar{\boldsymbol{A}}\right] - \mathbb{E}\left[\langle \widetilde{\boldsymbol{A}}, \boldsymbol{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}}\right] \right\|_{\text{op}}}_{\text{Term 3}} \\ & + \underbrace{\left\| \mathbb{E}\left[\langle \widetilde{\boldsymbol{A}}, \boldsymbol{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}}\right] - \frac{1}{n}\sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}}_i \right\|_{\text{op}}}_{\text{Term 4}} \\ & + \underbrace{\left\| \mathbb{E}\left[\bar{\eta}\bar{\boldsymbol{A}}\right] \right\|_{\text{op}}}_{\text{Term 5}}. \end{aligned} \tag{55}$$

In the remaining proof, we bound the five terms in (55) individually. We first discuss the nature of the five terms.

- **Terms 1 and 4:** These two terms characterize the difference between the empirical mean of quantities involving $\widetilde{\boldsymbol{A}}$ and their true expectation. In the proof, we show that the empirical mean concentrates around the expectation with high probability (see Lemma 5 and Lemma 8).
- **Terms 2 and 3:** These two terms characterize the difference in expectation introduced by truncating $\bar{\boldsymbol{A}}$ to $\widetilde{\boldsymbol{A}}$. Hence, these two terms characterize biases that arise from truncation. In the proof, these two terms diminish as $\tau \to \infty$ (see Lemma 6 and Lemma 7). Note that setting τ to ∞ is equivalent to no thresholding, and in this case $\widetilde{\boldsymbol{A}}$ becomes identical to $\bar{\boldsymbol{A}}$, and both terms diminish.
- **Term 5:** Term 5 is a bias term that arises from the fact that the mean of the noise η conditioned on sensing matrix $\bar{\boldsymbol{A}}$ is non-zero: $\mathbb{E}\left[\bar{\eta}|\bar{\boldsymbol{A}}\right] \neq 0$. We will show that this bias scales like $\frac{1}{m}$, allowing us to set the averaging number m to obtain consistent estimation.

By setting the truncation threshold τ carefully, we can make the Term 3 and 4 biases the same order as Terms 1 and 4.

Step 2: bound the five terms individually. In what follows, we provide five lemmas to bound each of the five terms individually. In the proofs of the five lemmas, we rely on an upper bound on the fourth moment of the m-sample averaged measurements $\bar{\gamma}^2$. Recall from Lemma 4 in Appendix A.5 that we have $\mathbb{E}[(\bar{\gamma}^2)^4] \leq M = c\left(\frac{y+\eta^\uparrow}{\sigma_r r}\right)^4$. We also rely heavily on the following truncation properties relating the m-sample averaged measurements $\bar{\gamma}^2$ and truncated measurements $\tilde{\gamma}^2$:

$$\tilde{\gamma}_i^2 \leq \tau \tag{TP1}$$
$$\tilde{\gamma}_i^2 \leq \bar{\gamma}_i^2 \tag{TP2}$$
$$\bar{\gamma}_i^2 - \tilde{\gamma}_i^2 = (\bar{\gamma}_i^2 - \tilde{\gamma}_i^2) \cdot \mathbf{1}\{\bar{\gamma}_i^2 \geq \tau\}. \tag{TP3}$$

The following lemma provides a bound for Term 1.

Lemma 5. *Let $\widetilde{\boldsymbol{A}}_1, \ldots, \widetilde{\boldsymbol{A}}_n$ be i.i.d copies of a random matrix $\widetilde{\boldsymbol{A}}$ as defined in Eq. (14). There exists a universal constant $c > 0$ such that the following is true. For any $t > 0$, we have*

$$\left\| \frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{A}}_i - \mathbb{E}\left[\widetilde{\boldsymbol{A}}_i\right] \right\|_{\text{op}} \lesssim \sqrt{\frac{M^{1/2}t}{n}} + \frac{\tau t}{n} \tag{56}$$

with probability at least $1 - 2 \cdot 9^d \cdot \exp(-t)$.

The proof of Lemma 5 is provided in Appendix C.1. The next lemma provides a deterministic upper bound for Term 2.

Lemma 6. *Let $\bar{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{A}}$ be the random matrices defined in Eq. (13) and Eq. (14), respectively. Then the bound*

$$\left\|\mathbb{E}\left[\widetilde{\boldsymbol{A}}\right]-\mathbb{E}\left[\bar{\boldsymbol{A}}\right]\right\|_{\mathrm{op}} \lesssim \frac{M^{1/2}}{\tau} \tag{57}$$

holds.

The proof of Lemma 6 is provided in Appendix C.2. The following lemma provides a deterministic upper bound for Term 3.

Lemma 7. *Let $\bar{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{A}}$ be the random matrices defined in Eq. (13) and Eq. (14), respectively. Then the bound*

$$\left\|\mathbb{E}\left[\langle\bar{\boldsymbol{A}},\boldsymbol{\Sigma}^{\star}\rangle\bar{\boldsymbol{A}}\right]-\mathbb{E}\left[\langle\widetilde{\boldsymbol{A}},\boldsymbol{\Sigma}^{\star}\rangle\widetilde{\boldsymbol{A}}\right]\right\|_{\mathrm{op}} \lesssim \frac{(y+\eta^{\uparrow})M^{1/2}}{\tau} \tag{58}$$

holds.

The proof of Lemma 7 is provided in Appendix C.3. The following lemma provides a bound for Term 4.

Lemma 8. *Let $\widetilde{\boldsymbol{A}}_1,\ldots,\widetilde{\boldsymbol{A}}_n$ be i.i.d copies of a random matrix $\widetilde{\boldsymbol{A}}$ as defined in Eq. (14). There exists a universal constant $c>0$ such that the following is true. For any $t>0$, we have*

$$\left\|\mathbb{E}\left[\langle\widetilde{\boldsymbol{A}},\boldsymbol{\Sigma}^{\star}\rangle\widetilde{\boldsymbol{A}}\right]-\frac{1}{n}\sum_{i=1}^{n}\langle\widetilde{\boldsymbol{A}}_i,\boldsymbol{\Sigma}^{\star}\rangle\widetilde{\boldsymbol{A}}_i\right\|_{\mathrm{op}} \lesssim (y+\eta^{\uparrow})\left(\sqrt{\frac{M^{1/2}t}{n}}+\frac{\tau t}{n}\right) \tag{59}$$

with probability at least $1-2\cdot 9^d\cdot\exp(-t)$.

The proof of Lemma 8 is provided in Appendix C.4. We note that Terms 2 and 3 are bias that result from shrinkage, but crucially are inversely dependent on the shrinkage threshold τ. This fact allows us to set τ so that the order of Terms 2 and 3 match the order of Terms 1 and 4. In particular, with the choice of $\tau = M^{1/4}\sqrt{\frac{n}{d}}$, all terms are of order $M^{1/4}\sqrt{\frac{d}{n}}$.

The final lemma bounds Term 5, which is a bias that arises from the dependence of the sensing matrix $\bar{\boldsymbol{A}}$ on the noise η.

Lemma 9. *Let $\bar{\boldsymbol{A}}$ be the random matrix defined in Eq. (13). Suppose that $\boldsymbol{\Sigma}^{\star}$ has rank r with $r>2$. Then we have*

$$\mathbb{E}\left[\left\|\bar{\eta}\bar{\boldsymbol{A}}\right\|_{\mathrm{op}}\right] \lesssim \frac{1}{\sigma_r r}\frac{\nu_\eta^2}{m}. \tag{60}$$

The proof of Lemma 9 is provided in Appendix C.5. We note that the bias scales with the variance of the m-sample averaged noise $\bar{\eta}$, which scales inversely with m.

Step 3: combine the five terms. We set $t=(\log 9+1)d$ and denote $y^{\uparrow}=y+\eta^{\uparrow}$. Utilizing Lemmas 5– 9, we arrive at an upper bound for the operator norm. We have that with probability at least $1-4\exp(-d)$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}y\widetilde{\boldsymbol{A}}_i-\frac{1}{n}\sum_{i=1}^{n}\langle\widetilde{\boldsymbol{A}}_i,\boldsymbol{\Sigma}^{\star}\rangle\widetilde{\boldsymbol{A}}_i\right\|_{\mathrm{op}} \lesssim (y^{\uparrow}+1)\left(\sqrt{\frac{M^{1/2}d}{n}}+\frac{d}{n}\tau+\frac{M^{1/2}}{\tau}\right)+\frac{1}{\sigma_r r}\frac{\nu_\eta^2}{m} \tag{61}$$

$$\overset{(i)}{\lesssim} y^{\uparrow}\left(\frac{y^{\uparrow}}{\sigma_r r}\sqrt{\frac{d}{n}}+\frac{d}{n}\tau+\left(\frac{y^{\uparrow}}{\sigma_r r}\right)^2\frac{1}{\tau}+\frac{1}{\sigma_r r}\frac{\nu_\eta^2}{m}\right) \tag{62}$$

as desired. Above, (i) follows from substituting in the expression for M from Lemma 4.

B.2. Proof of Proposition 3

Our objective is to show that the there exists some constant κ such that the RSC condition

$$\frac{1}{n}\sum_{i=1}^{n}\langle\widetilde{\boldsymbol{A}}_i,\boldsymbol{U}\rangle^2 \geq \kappa\|\boldsymbol{U}\|_F^2 \tag{63}$$

holds uniformly for all matrices $\boldsymbol{U}$ in the error set

$$\mathcal{E} = \left\{ \boldsymbol{U} \in \mathbb{S}^{d\times d} : \|\boldsymbol{U}\|_* \le 4\sqrt{2r}\|\boldsymbol{U}\|_F \right\}. \tag{64}$$

Recall from the definition of $\widetilde{\boldsymbol{A}}$ that

$$\widetilde{\boldsymbol{A}}_i = \widetilde{\gamma}_i^2 \boldsymbol{a}_i \boldsymbol{a}_i^\top \tag{65}$$

$$= \left(\frac{y + \bar{\eta}_i}{\boldsymbol{a}_i^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}_i} \wedge \tau \right) \boldsymbol{a}_i \boldsymbol{a}_i^\top \tag{66}$$

so we have

$$\langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{U} \rangle^2 = \left(\frac{y + \bar{\eta}_i}{\boldsymbol{a}_i^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}_i} \wedge \tau \right)^4 \left(\boldsymbol{a}_i^\top \boldsymbol{U} \boldsymbol{a}_i \right)^2.$$

This implies that $\sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{U} \rangle^2$ is nondecreasing in τ when $\tau > 0$. As a result, defining the random matrix

$$\widetilde{\boldsymbol{A}}^{\tau'} = \left(\frac{y + \bar{\eta}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \wedge \tau' \right) \boldsymbol{a}\boldsymbol{a}^\top, \tag{67}$$

we have that the following lower bound holds for any $\tau' \le \tau$.

$$\frac{1}{n} \sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{U} \rangle^2 \ge \frac{1}{n} \sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i^{\tau'}, \boldsymbol{U} \rangle^2. \tag{68}$$

Above $\widetilde{\boldsymbol{A}}_1^{\tau'}, \ldots, \widetilde{\boldsymbol{A}}_n^{\tau'}$ are i.i.d copies of the random matrix (67). We will lower bound $\frac{1}{n} \sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i^{\tau'}, \boldsymbol{U} \rangle^2$ for an appropriate value of τ', which we will set later. To proceed, we will use a small-ball argument (Mendelson, 2015; Tropp, 2015) based on the following lemma.

Lemma 10 (Tropp, 2015, Proposition 5.1, adapted to our notation)**.** *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^{d\times d}$ be i.i.d. copies of a random matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$. Let $E \subset \mathbb{R}^{d\times d}$. Let $\xi, Q > 0$ be such that for every $\boldsymbol{U} \in E$,*

$$\mathbb{P}\left(|\langle \boldsymbol{X}, \boldsymbol{U} \rangle| \ge 2\xi \right) \ge Q. \tag{69}$$

Furthermore, denote the Rademacher width as

$$W = \mathbb{E}\left[\sup_{\boldsymbol{U}\in E} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle \boldsymbol{X}_i, \boldsymbol{U} \rangle \right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables independent of the $\boldsymbol{X}_i$'s. Then, for any $t > 0$, with probability at least $1 - \exp\left(-\frac{nt^2}{2}\right)$,

$$\inf_{\boldsymbol{U}\in E} \left(\frac{1}{n} \sum_{i=1}^n \langle \boldsymbol{X}_i, \boldsymbol{U} \rangle^2 \right)^{1/2} \ge \xi(Q - t) - 2W.$$

We apply Lemma 10 with $\boldsymbol{X}_i = \widetilde{\boldsymbol{A}}_i^{\tau'}$ and with set E as

$$E = \mathcal{E} \cap \{ \boldsymbol{U} \in \mathbb{R}^{d\times d} : \|\boldsymbol{U}\|_F = 1 \} \tag{70}$$

$$= \{ \boldsymbol{U} \in \mathbb{S}^{d\times d} : \|\boldsymbol{U}\|_F = 1, \|\boldsymbol{U}\|_* \le 4\sqrt{2r} \} \tag{71}$$

The rest of the proof is comprised of two key steps. To invoke Lemma 10, the first step establishes the inequality (69) by lower bounding Q. The second step upper bounds the Rademacher width W. The following lemma provides the lower bound on Q.

Lemma 11. *Consider any $\tau' \in (0, \tau)$. There exist absolute constants $c_1, c_2 > 0$ such that for every $\boldsymbol{U} \in E$, we have*

$$\mathbb{P}\left(\left|\langle \widetilde{\boldsymbol{A}}^{\tau'}, \boldsymbol{U}\rangle\right| \geq c_1 \left(\frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)} \wedge \tau'\right)\right) \geq c_2. \tag{72}$$

The proof of Lemma 11 is presented in Appendix D.1. We now turn to the second step of the proof, which is bounding the Rademacher width W. The next lemma characterizes this width.

Lemma 12. *Consider any $\tau' \in (0, \tau)$. Let $\widetilde{\boldsymbol{A}}_1^{\tau'}, \ldots, \widetilde{\boldsymbol{A}}_n^{\tau'} \in \mathbb{R}^{d\times d}$ be i.i.d. copies of the random matrix $\widetilde{\boldsymbol{A}}^{\tau'} \in \mathbb{R}^{d\times d}$ defined in Equation (67). Let E be the set defined in Equation (71). Then, there exists some absolute constants c_1 and c_2 such that if*

$$n \geq c_1 d \tag{73a}$$

the bound

$$\mathbb{E}\left[\sup_{\boldsymbol{U}\in E} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \widetilde{\boldsymbol{A}}_i^{\tau'}, \boldsymbol{U}\rangle\right] \leq c_2 \tau' \sqrt{\frac{rd}{n}} \tag{73b}$$

holds.

The proof of Lemma 12 is presented in Appendix D.2. Invoking Lemma 11 and Lemma 12, we have that for some constant c_4, as long as $n \geq c_4 d$, the bound

$$\begin{aligned}\inf_{\boldsymbol{U}\in E}\left(\frac{1}{n}\sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{U}\rangle^2\right)^{1/2} &\geq \inf_{\boldsymbol{U}\in E}\left(\frac{1}{n}\sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i^{\tau'}, \boldsymbol{U}\rangle^2\right)^{1/2} \\ &\geq c_1' \left(\frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)} \wedge \tau'\right)(c_2 - t) - c_3 \tau' \sqrt{\frac{rd}{n}}\end{aligned} \tag{74}$$

with probability at least $1 - \exp\left(-\frac{nt^2}{2}\right)$. We set $\tau' = \frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)}$, where κ_y is the median of the random quantity $y + \bar{\eta}$. By the assumption $\tau \geq \frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)}$, this choice of τ' satisfies $\tau' \leq \tau$. Setting $t = \frac{c_2}{2}$, we have for some constant c, that with probability at least $1 - \exp(-cn)$,

$$\inf_{\boldsymbol{U}\in E} \frac{1}{n}\left(\sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i^{\tau'}, \boldsymbol{U}\rangle^2\right)^{1/2} \geq \frac{c_1' c_2}{2}\frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)} - c_3 \frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)}\sqrt{\frac{rd}{n}}. \tag{75}$$

Therefore, if $n \geq \max\left\{\left(\frac{4c_3}{c_1' c_2}\right)^2, c_4\right\} rd$, we have

$$\inf_{\boldsymbol{U}\in\mathcal{E}} \frac{1}{n}\sum_{i=1}^n \langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{U}\rangle^2 \geq \left(\frac{c_1' c_2}{4}\frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)}\right)^2 \|\boldsymbol{U}\|_F^2 \tag{76}$$

with probability at least $1 - \exp(-cn)$. We conclude by setting $\kappa_{\mathcal{L}} = \left(\frac{c_1' c_2}{4}\right)^2$ and $C = \max\left\{\left(\frac{4c_3}{c_1' c_2}\right)^2, c_4\right\}$.

C. Proof of supporting lemmas for Proposition 2

In this section, we prove the supporting lemmas for Proposition 2.

C.1. Proof of Lemma 5.

Let $\mathcal{S}_{\frac{1}{4}} \subseteq \mathcal{S}^{d-1}$ be a $\frac{1}{4}$-covering of unit-norm d-dimensional vectors. By a covering argument (Vershynin, 2018, Exercise 4.4.3), for any symmetric matrix $\boldsymbol{U} \in \mathbb{R}^{d\times d}$, its operator norm is bounded by $\|\boldsymbol{U}\|_{\text{op}} \leq 2\sup_{\boldsymbol{v}\in\mathcal{S}_{\frac{1}{4}}} \left|\boldsymbol{v}^\top \boldsymbol{U} \boldsymbol{v}\right|$. Hence, we

have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{A}}_i-\mathbb{E}\left[\widetilde{\boldsymbol{A}}\right]\right\|_{\mathrm{op}}\leq 2\sup_{\boldsymbol{v}\in\mathcal{S}_{\frac{1}{4}}}\left|\boldsymbol{v}^\top\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{A}}_i-\mathbb{E}\left[\widetilde{\boldsymbol{A}}\right]\right)\boldsymbol{v}\right|$$
$$=2\sup_{\boldsymbol{v}\in\mathcal{S}_{\frac{1}{4}}}\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^\top\widetilde{\boldsymbol{A}}_i\boldsymbol{v}-\mathbb{E}\left[\boldsymbol{v}^\top\widetilde{\boldsymbol{A}}\boldsymbol{v}\right]\right|. \tag{77}$$

We now apply Bernstein's inequality to bound (77). We first assume the Bernstein condition holds with $u_1=c_1M^{\frac{1}{2}}$ and $u_2=c_2\tau$ for some universal positive constants c_1,c_2. Namely, for each integer $p\geq 2$, we show that for any unit vector $\boldsymbol{v}\in\mathbb{R}^d$,

$$\mathbb{E}\left[\left|\boldsymbol{v}^\top\widetilde{\boldsymbol{A}}\boldsymbol{v}\right|^p\right]\leq\frac{p!}{2}u_1u_2^{p-2}. \tag{78}$$

We first provide the rest of the proof assuming that (78) holds, followed by proving (78). By Bernstein's inequality (see Lemma 2), under condition (78), we have that for any unit vector $\boldsymbol{v}\in\mathbb{R}^d$ and any $t>0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^\top\widetilde{\boldsymbol{A}}_i\boldsymbol{v}-\mathbb{E}\left[\boldsymbol{v}^\top\widetilde{\boldsymbol{A}}\boldsymbol{v}\right]\right|\geq 2\left(\sqrt{\frac{u_1M^{1/2}t}{n}}+\frac{u_2\tau t}{n}\right)\right)\leq 2\exp\left(-t\right). \tag{79}$$

By Vershynin (2018, Corollary 4.2.13), the cardinality of the covering set $\mathcal{S}_{\frac{1}{4}}$ is bounded above by 9^d. Therefore, by a union bound,

$$\mathbb{P}\left(\sup_{\boldsymbol{v}\in\mathcal{S}_{\frac{1}{4}}}\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^\top\widetilde{\boldsymbol{A}}_i\boldsymbol{v}-\mathbb{E}\left[\boldsymbol{v}^\top\widetilde{\boldsymbol{A}}\boldsymbol{v}\right]\right|\geq 2\left(\sqrt{\frac{u_1M^{1/2}t}{n}}+\frac{u_2\tau t}{n}\right)\right)\leq 2\cdot 9^d\cdot\exp\left(-t\right). \tag{80}$$

Combining (77) and (80), for any $t>0$, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{A}}_i-\mathbb{E}\left[\widetilde{\boldsymbol{A}}\right]\right\|_{\mathrm{op}}\lesssim\sqrt{\frac{M^{1/2}t}{n}}+\frac{\tau t}{n} \tag{81}$$

with probability at least $1-2\cdot 9^d\cdot\exp\left(-t\right)$, as desired. It remains to prove the Bernstein condition (78).

Proving the Bernstein condition (78) holds. We fix any unit vector $\boldsymbol{v}\in\mathbb{R}^d$. Plugging in $\widetilde{\boldsymbol{A}}=\widetilde{\gamma}^2\boldsymbol{a}\boldsymbol{a}^\top$, we have $\boldsymbol{v}^\top\widetilde{\boldsymbol{A}}\boldsymbol{v}=\widetilde{\gamma}^2(\boldsymbol{v}^\top\boldsymbol{a})^2$. Since the random variable $\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)$, and $\boldsymbol{v}$ is a unit vector, it follows that $\boldsymbol{v}^\top\boldsymbol{a}\sim\mathcal{N}(0,1)$. Denote by $G\sim\mathcal{N}(0,1)$ a standard normal random variable. For any integer $p\geq 2$, we have

$$\begin{aligned}\mathbb{E}[\left|\boldsymbol{v}^\top\widetilde{\boldsymbol{A}}\boldsymbol{v}\right|^p]=\mathbb{E}\left[(\widetilde{\gamma}^2G^2)^p\right]&\overset{\text{(i)}}{\leq}\tau^{p-2}\mathbb{E}\left[\left(\widetilde{\gamma}^2\right)^2G^{2p}\right]\\&\overset{\text{(ii)}}{\leq}\tau^{p-2}\cdot\mathbb{E}\left[\left(\bar{\gamma}^2\right)^2G^{2p}\right]\\&\overset{\text{(iii)}}{\leq}\tau^{p-2}\left(\mathbb{E}\left[\left(\bar{\gamma}^2\right)^4\right]\cdot\mathbb{E}\left[G^{4p}\right]\right)^{1/2}\\&\overset{\text{(iv)}}{\leq}\tau^{p-2}\left(M\cdot\mathbb{E}\left[G^{4p}\right]\right)^{1/2},\end{aligned} \tag{82}$$

where steps (i) and (ii) follow from TP1 and TP2, respectively, step (iii) follows from Cauchy–Schwarz, and step (iv) follows from the upper bound on the fourth moment of $\bar{\gamma}^2$. Note that G^2 follows a Chi-Square distribution with 1 degree of freedom, and hence sub-exponential. Recall from (17) in Appendix A.2 that there exists some constant $c>0$ such that we have $\left(\mathbb{E}\left[(G^2)^p\right]\right)^{1/p}\leq cp$ for all $p\geq 1$. Hence, we have

$$\begin{aligned}\left(\mathbb{E}\left[G^{4p}\right]\right)^{1/2}&\leq(2cp)^p=(2ec)^p\cdot(\frac{p}{e})^p\\&\overset{\text{(i)}}{<}p!\cdot(2ec)^p\end{aligned} \tag{83}$$

where step (i) is true by Stirling's inequality that for all $p \geq 1$,

$$p! > \sqrt{2\pi p}\left(\frac{p}{e}\right)^p e^{\frac{1}{12p+1}} > (\frac{p}{e})^p.$$

Plugging (83) to (82) and rearranging terms completes the proof of Bernstein condition (78).

C.2. Proof of Lemma 6.

We first begin by showing that $\mathbb{E}\left[\bar{\boldsymbol{A}}\right] - \mathbb{E}\left[\widetilde{\boldsymbol{A}}\right] \succeq \mathbf{0}$. Substituting in the definitions of $\bar{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{A}}$, we have $\mathbb{E}\left[\bar{\boldsymbol{A}}\right] - \mathbb{E}\left[\widetilde{\boldsymbol{A}}\right] = \mathbb{E}\left[(\bar{\gamma}^2 - \widetilde{\gamma}^2)\boldsymbol{a}\boldsymbol{a}^\top\right]$. By TP2, we have $\bar{\gamma}^2 \geq \widetilde{\gamma}^2$, meaning that $\bar{\gamma}^2 - \widetilde{\gamma}^2$ is non-negative. The expectation of a non-negative quantity times an outer product is symmetric positive semidefinite. Therefore, we can write the operator norm as

$$\left\|\mathbb{E}\left[\widetilde{\boldsymbol{A}}\right] - \mathbb{E}\left[\bar{\boldsymbol{A}}\right]\right\|_{\text{op}} = \sup_{\boldsymbol{v}\in\mathcal{S}^{d-1}} \boldsymbol{v}^\top\left(\mathbb{E}\left[\bar{\boldsymbol{A}}\right] - \mathbb{E}\left[\widetilde{\boldsymbol{A}}\right]\right)\boldsymbol{v}. \tag{84}$$

We now show that there exists a uniform upper bound on the quantity

$$\boldsymbol{v}^\top\left(\mathbb{E}\left[\bar{\boldsymbol{A}}\right] - \mathbb{E}\left[\widetilde{\boldsymbol{A}}\right]\right)\boldsymbol{v} \tag{85}$$

for a unit-norm d-dimensional vector $\boldsymbol{v}$, therefore bounding the operator norm. We again note $(\boldsymbol{v}^\top\boldsymbol{a}) \sim \mathcal{N}(0,1)$ and denote $G \sim \mathcal{N}(0,1)$. Then we rewrite the quantity (85) by substituting in the expression for sensing matrices $\bar{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{A}}$, as follows.

$$\begin{aligned}
\boldsymbol{v}^\top\left(\mathbb{E}\left[\bar{\boldsymbol{A}}\right] - \mathbb{E}\left[\widetilde{\boldsymbol{A}}\right]\right)\boldsymbol{v} &= \mathbb{E}\left[\boldsymbol{v}^\top\left(\bar{\boldsymbol{A}} - \widetilde{\boldsymbol{A}}\right)\boldsymbol{v}\right] && (86)\\
&= \mathbb{E}\left[\boldsymbol{v}^\top\left(\bar{\gamma}^2\boldsymbol{a}\boldsymbol{a}^\top - \widetilde{\gamma}^2\boldsymbol{a}\boldsymbol{a}^\top\right)\boldsymbol{v}\right] && (87)\\
&= \mathbb{E}\left[\left(\bar{\gamma}^2 - \widetilde{\gamma}^2\right)G^2\right]. && (88)
\end{aligned}$$

We continue from here by utilizing properties of the shrunken measurements, as follows.

$$\begin{aligned}
\mathbb{E}\left[\left(\bar{\gamma}^2 - \widetilde{\gamma}^2\right)G^2\right] &\overset{\text{(i)}}{=} y\,\mathbb{E}\left[\left(\bar{\gamma}^2 - \widetilde{\gamma}^2\right)G^2\mathbf{1}\{\bar{\gamma}^2 \geq \tau\}\right] && (89)\\
&\leq \mathbb{E}\left[\bar{\gamma}^2 G^2\mathbf{1}\{\bar{\gamma}^2 \geq \tau\}\right] && (90)\\
&\overset{\text{(ii)}}{\leq} \left(\mathbb{E}\left[(\bar{\gamma}^2G^2)^2\right]\cdot\mathbb{E}\left[\mathbf{1}\{\bar{\gamma}^2 \geq \tau\}\right]\right)^{1/2} && (91)\\
&\overset{\text{(iii)}}{\leq} \left(\mathbb{E}\left[|\bar{\gamma}^2|^4\right]\cdot\mathbb{E}\left[|G^2|^4\right]\right)^{1/4}\left(\mathbb{P}\left(\bar{\gamma}^2 \geq \tau\right)\right)^{1/2}, && (92)
\end{aligned}$$

where step (i) follows from TP3, and (ii) and (iii) follow from Cauchy–Schwarz. We proceed by bounding each of the above terms separately. First, recall from Lemma 4 in Appendix A.5 that the fourth moment $\mathbb{E}\left[|\bar{\gamma}^2|^4\right]$ is bounded above by M. Second, G^2 is a sub-exponential random variable. By Appendix A.2, we have that $\mathbb{E}\left[|G^2|^4\right]^{1/4} \leq c$ for some constant c. It remains to bound $\left(\mathbb{P}\left(\bar{\gamma}^2 \geq \tau\right)\right)^{1/2}$, which we do below.

$$\begin{aligned}
\left(\mathbb{P}\left(\bar{\gamma}^2 \geq \tau\right)\right)^{1/2} &\overset{(iv)}{\leq} \left(\frac{\mathbb{E}\left[|\bar{\gamma}^2|^2\right]}{\tau^2}\right)^{1/2} && (93)\\
&\overset{(v)}{\leq} \frac{\left(\mathbb{E}\left[|\bar{\gamma}^2|^4\right]\right)^{1/4}}{\tau} && (94)\\
&\overset{(vi)}{\leq} \frac{M^{1/4}}{\tau}. && (95)
\end{aligned}$$

Above, (iv) follows from Markov's inequality, (v) follows from Cauchy–Schwarz, and (vi) follows from the fourth moment bound on the averaged scaling $\bar{\gamma}^2$. Putting everything together, we have that the bound

$$\boldsymbol{v}^\top\left(\mathbb{E}\left[\bar{\boldsymbol{A}}\right] - \mathbb{E}\left[\widetilde{\boldsymbol{A}}\right]\right)\boldsymbol{v} \lesssim \frac{M^{1/2}}{\tau} \tag{96}$$

holds uniformly for all vectors $\boldsymbol{v} \in \mathcal{S}^{d-1}$. Therefore,

$$\left\| \mathbb{E}\left[\widetilde{\boldsymbol{A}}\right] - \mathbb{E}\left[\bar{\boldsymbol{A}}\right] \right\|_{\text{op}} \lesssim \frac{M^{1/2}}{\tau}, \tag{97}$$

as desired

C.3. Proof of Lemma 7.

We begin by substituting in the definition $\bar{\boldsymbol{A}} = \bar{\gamma}^2 \boldsymbol{a}\boldsymbol{a}^\top$, the matrix $\langle \bar{\boldsymbol{A}}, \boldsymbol{\Sigma}^\star \rangle \bar{\boldsymbol{A}}$ can be written as $\bar{\gamma}^4 \boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}$. Similarly, we can re-write the matrix $\langle \widetilde{\boldsymbol{A}}, \boldsymbol{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}}$ as $\widetilde{\gamma}^4 \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) \boldsymbol{a}\boldsymbol{a}^\top$. Therefore, our goal is to bound the operator norm

$$\left\| \left(\bar{\gamma}^4 - \widetilde{\gamma}^4\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) \boldsymbol{a}\boldsymbol{a}^\top \right\|_{\text{op}} \tag{98}$$

We note that the matrix $\left(\bar{\gamma}^4 - \widetilde{\gamma}^4\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) \boldsymbol{a}\boldsymbol{a}^\top$ is symmetric positive semidefinite, as it is the product of a non-negative scalar $\left(\bar{\gamma}^4 - \widetilde{\gamma}^4\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right)$ and an outer product. Similar to the proof of Lemma 6, we now show a uniform upper bound on the quantity $\boldsymbol{v}^\top \mathbb{E}\left[\left(\bar{\gamma}^4 - \widetilde{\gamma}^4\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) \boldsymbol{a}\boldsymbol{a}^\top\right] \boldsymbol{v}$ for arbitrary unit-norm d-dimensional vector $\boldsymbol{v}$.

Again, note that $\boldsymbol{v}^\top \boldsymbol{a} \sim \mathcal{N}(0,1)$ and denote $G \sim \mathcal{N}(0,1)$. We begin by substituting in the expressions for G.

$$\boldsymbol{v}^\top \mathbb{E}\left[\left(\bar{\gamma}^4 - \widetilde{\gamma}^4\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) \boldsymbol{a}\boldsymbol{a}^\top\right] \boldsymbol{v} = \mathbb{E}\left[\left(\bar{\gamma}^4 - \widetilde{\gamma}^4\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) \boldsymbol{v}^\top \boldsymbol{a}\boldsymbol{a}^\top \boldsymbol{v}\right] \tag{99}$$

$$= \mathbb{E}\left[\left(\bar{\gamma}^4 - \widetilde{\gamma}^4\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) G^2\right] \tag{100}$$

Next, we can proceed by manipulating the $\bar{\gamma}^4 - \widetilde{\gamma}^4$ term to remove the term $\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}$, as follows.

$$\mathbb{E}\left[\left(\bar{\gamma}^4 - \widetilde{\gamma}^4\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) G^2 \mathbf{1}\{\gamma^2 \geq \tau\}\right] = \mathbb{E}\left[\left(\bar{\gamma}^2 + \widetilde{\gamma}^2\right) \left(\bar{\gamma}^2 - \widetilde{\gamma}^2\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) G^2\right] \tag{101}$$

$$\overset{(i)}{\leq} \mathbb{E}\left[2\bar{\gamma}^2 \left(\bar{\gamma}^2 - \widetilde{\gamma}^2\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) G^2\right] \tag{102}$$

$$\overset{(ii)}{=} 2\mathbb{E}\left[(y + \bar{\eta}) \left(\bar{\gamma}^2 - \widetilde{\gamma}^2\right) G^2\right] \tag{103}$$

$$\overset{(iii)}{\leq} 2(y + \eta^\uparrow)\mathbb{E}\left[\left(\bar{\gamma}^2 - \widetilde{\gamma}^2\right) G^2 \mathbf{1}\{\gamma^2 \geq \tau\}\right] \tag{104}$$

Above, (i) follows from TP2, (ii) follows from the definition $\bar{\gamma}^2 = \frac{y + \bar{\eta}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}$, and (iii) follows from TP3 and the upper bound on noise η.

The rest of the proof follows the exact steps of the proof of Lemma 6, provided in Section C.2. Therefore, we have the bound

$$\left\| \mathbb{E}\left[\left(\bar{\gamma}^4 - \widetilde{\gamma}^4\right) \left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right) \boldsymbol{a}\boldsymbol{a}^\top\right] \right\|_{\text{op}} \lesssim \frac{(y + \eta^\uparrow) M^{1/2}}{\tau}, \tag{105}$$

as desired.

C.4. Proof of Lemma 8

The proof follows the steps as the proof of Lemma 5, and we explain the difference where we now provide a Bernstein condition with $u_1 = c_1(y + \eta^\uparrow)^2$ and $u_2 = c_2(y + \eta^\uparrow)\tau$. Namely, for every integer $p \geq 2$, we have (cf. 78)

$$\mathbb{E}\left[\left|\boldsymbol{v}^\top \langle \widetilde{\boldsymbol{A}}, \boldsymbol{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}} \boldsymbol{v}\right|^p\right] \leq \frac{p!}{2} u_1 u_2^{p-2}. \tag{106}$$

Plugging in $\widetilde{\boldsymbol{A}} = \widetilde{\gamma}^2 \boldsymbol{a}\boldsymbol{a}^\top$, we have

$$\mathbb{E}\left[\left|\boldsymbol{v}^\top \langle \widetilde{\boldsymbol{A}}, \boldsymbol{\Sigma}^\star \rangle \widetilde{\boldsymbol{A}} \boldsymbol{v}\right|^p\right] = \mathbb{E}\left[\left(\widetilde{\gamma}^2 \boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right)^p \cdot \left|\boldsymbol{v}^\top \widetilde{\boldsymbol{A}} \boldsymbol{v}\right|^p\right]$$

$$\overset{\text{(i)}}{\leq} \mathbb{E}\left[\left(\bar{\gamma}^2 \boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}\right)^p \cdot \left|\boldsymbol{v}^\top \widetilde{\boldsymbol{A}} \boldsymbol{v}\right|^p\right] \tag{107}$$

$$\overset{\text{(ii)}}{=} \mathbb{E}\left[(y + \bar{\eta})^p \cdot \left|\boldsymbol{v}^\top \widetilde{\boldsymbol{A}} \boldsymbol{v}\right|^p\right] \tag{108}$$

$$\overset{\text{(iii)}}{\leq} (y + \eta^\uparrow)^p \cdot \mathbb{E}\left[\left|\boldsymbol{v}^\top \widetilde{\boldsymbol{A}} \boldsymbol{v}\right|^p\right]. \tag{109}$$

Plugging in (78) from Lemma 5 to bound the term $\mathbb{E}\left[\left|\boldsymbol{v}^\top \widetilde{\boldsymbol{A}}\boldsymbol{v}\right|^p\right]$ in (109) completes the proof of the Bernstein condition (106).

Above, (i) follows from TP2, (ii) follows from the definition $\bar{\gamma}^2 = \frac{y+\bar{\eta}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}$, and (iii) follows from the upper bound on the noise η. The rest of the proof follows in the same manner as the proof of Lemma 5, as presented in Section C.1, with an additional factor of $y + \eta^\uparrow$. Therefore, like in Section C.1, the bound

$$\left\|\mathbb{E}\left[\langle \widetilde{\boldsymbol{A}}, \boldsymbol{\Sigma}^\star\rangle \widetilde{\boldsymbol{A}}\right] - \frac{1}{n}\sum_{i=1}^{n}\langle \widetilde{\boldsymbol{A}}_i, \boldsymbol{\Sigma}^\star\rangle \widetilde{\boldsymbol{A}}_i\right\|_{\text{op}} \lesssim (y+\eta^\uparrow)\left(\sqrt{\frac{M^{1/2}t}{n}} + \frac{\tau t}{n}\right) \tag{110}$$

holds with probability greater than $1 - 2\cdot 9^d \cdot \exp(-t)$, as desired.

C.5. Proof of Lemma 9.

Substituting in $\bar{\boldsymbol{A}} = \bar{\gamma}^2 \boldsymbol{a}\boldsymbol{a}^\top = \frac{y+\bar{\eta}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \boldsymbol{a}\boldsymbol{a}^\top$, we have

$$\begin{aligned}
\left\|\mathbb{E}\left[\bar{\eta}\bar{\boldsymbol{A}}\right]\right\|_{\text{op}} &= \left\|\mathbb{E}\left[\bar{\eta}(y+\bar{\eta})\frac{\boldsymbol{a}\boldsymbol{a}^\top}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right]\right\|_{\text{op}} \\
&= \left\|\mathbb{E}\left[\bar{\eta}(y+\bar{\eta})\right]\cdot \mathbb{E}\left[\frac{\boldsymbol{a}\boldsymbol{a}^\top}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right]\right\|_{\text{op}} \\
&= \frac{\sigma_\eta^2}{m}\left\|\mathbb{E}\left[\frac{\boldsymbol{a}\boldsymbol{a}^\top}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right]\right\|_{\text{op}}.
\end{aligned} \tag{111}$$

To bound the operator norm term in (111), recall from Lemma 3(b) in Appendix A.4 that for any matrix $\boldsymbol{U}$, we have

$$\mathbb{E}\left(\frac{\boldsymbol{a}^\top \boldsymbol{U}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right) \lesssim \frac{1}{\sigma_r r}\|\boldsymbol{U}\|_*. \tag{112}$$

Note that $\frac{\boldsymbol{a}\boldsymbol{a}^\top}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}$ is symmetric positive semidefinite, so we have

$$\begin{aligned}
\left\|\mathbb{E}\left[\frac{\boldsymbol{a}\boldsymbol{a}^\top}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right]\right\|_{\text{op}} &= \sup_{\boldsymbol{v}\in\mathcal{S}^{d-1}}\left|\boldsymbol{v}^\top \mathbb{E}\left[\frac{\boldsymbol{a}\boldsymbol{a}^\top}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right]\boldsymbol{v}\right| \\
&= \sup_{\boldsymbol{v}\in\mathcal{S}^{d-1}} \mathbb{E}\left[\frac{\boldsymbol{a}^\top(\boldsymbol{v}\boldsymbol{v}^\top)\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}}\right] \\
&\overset{\text{(i)}}{\lesssim} \frac{1}{\sigma_r r}\sup_{\boldsymbol{v}\in\mathcal{S}^{d-1}}\|\boldsymbol{v}\boldsymbol{v}^\top\|_* \\
&\overset{\text{(ii)}}{=} \frac{1}{\sigma_r r},
\end{aligned} \tag{113}$$

where step (i) is true by plugging in (112), and step (ii) is true because $\|\boldsymbol{v}\boldsymbol{v}^\top\|_* = 1$ for any unit norm vector $\boldsymbol{v}$. Plugging (113) back to (111), we have

$$\left\|\mathbb{E}\left[\bar{\eta}\bar{\boldsymbol{A}}\right]\right\|_{\text{op}} \lesssim \frac{1}{\sigma_r r}\cdot\frac{\nu_\eta^2}{m}, \tag{114}$$

as desired.

D. Proof of supporting lemmas for Proposition 3

In this section, we prove the supporting lemmas for Proposition 3.

D.1. Proof of Lemma 11

For the proof, we first fix any $\boldsymbol{U} \in \mathcal{E} \cap \{\boldsymbol{U} \in \mathbb{S}^{d\times d} : \|\boldsymbol{U}\|_F = 1\}$. Let κ_y be the median of $y + \bar{\eta}$ and let $\mathcal{G}$ be the event that $y + \eta \geq \kappa_y$, which occurs with probability $\frac{1}{2}$. For any $\xi > 0$, because the averaged noise $\bar{\eta}$ and sensing vector $\boldsymbol{a}$ are independent,

$$\mathbb{P}\left(\left|\langle \widetilde{\boldsymbol{A}}^{\tau'}, \boldsymbol{U}\rangle\right| \geq \xi\right) = \mathbb{P}\left(\left(\frac{y+\eta}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \wedge \tau'\right) \left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right| \geq \xi\right) \tag{115}$$

$$= \mathbb{P}\left(\left(\frac{y+\eta}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \wedge \tau'\right) \left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right| \geq \xi \middle| \mathcal{G}\right) \mathbb{P}(\mathcal{G}) \tag{116}$$

$$= \frac{1}{2}\mathbb{P}\left(\left(\frac{y+\eta}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \wedge \tau'\right) \left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right| \geq \xi \middle| \mathcal{G}\right) \tag{117}$$

$$\geq \frac{1}{2}\mathbb{P}\left(\left(\frac{\kappa_y}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \wedge \tau'\right) \left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right| \geq \xi\right) \tag{118}$$

We proceed by bounding the terms in (118) separately.

Lower bound on $\mathbb{P}\left(\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right| \geq c_1\right)$. We use the approach from Kueng et al. (2017, Section 4.1). By Paley-Zygmund inequality,

$$\mathbb{P}\left(\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right|^2 \geq \frac{1}{2}\mathbb{E}\left[\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right|^2\right]\right) \geq \frac{1}{4}\frac{\left(\mathbb{E}\left[\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right|^2\right]\right)^2}{\mathbb{E}\left[\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right|^4\right]} \tag{119}$$

As noted in Kueng et al. (2017, Section 4.1), there exists some constant c_2' such that for any matrix $\boldsymbol{U}$ with unit Frobenius norm,

$$\mathbb{E}\left[\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right|^2\right] \geq 1 \quad \text{and} \quad \mathbb{E}\left[\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right|^4\right] \leq c_2'\left(\mathbb{E}\left[\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right|^2\right]\right)^2. \tag{120}$$

Note that by the definition, every matrix $\boldsymbol{U} \in E$ has unit Frobenius norm. Utilizing Paley-Zygmund (119) and the bounds on the second and fourth moment of $\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle$ (120), there exist positive constants c_1 and c_2 such that

$$\mathbb{P}\left(\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right| \geq c_1\right) \geq c_2. \tag{121}$$

Upper bound on $\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}$. By Hanson-Wright inequality (Rudelson & Vershynin, 2013, Theorem 1.1), there exist some positive absolute constants c and c_3' such that for any $t > 0$, we have

$$\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a} \leq c_3'\left(\operatorname{tr}(\boldsymbol{\Sigma}^\star) + \|\boldsymbol{\Sigma}^\star\|_F \sqrt{t} + \|\boldsymbol{\Sigma}^\star\|_{\text{op}}\, t\right) \tag{122}$$

with probability at least $1 - 2\exp(-ct)$. Set t to be a constant such that $2\exp(-ct) = \frac{c_2}{2}$ and note that for symmetric positive semidefinite matrix $\boldsymbol{\Sigma}^\star$, the bounds $\|\boldsymbol{\Sigma}^\star\|_F \leq \operatorname{tr}(\boldsymbol{\Sigma}^\star)$ and $\|\boldsymbol{\Sigma}^\star\|_{\text{op}} \leq \operatorname{tr}(\boldsymbol{\Sigma}^\star)$ hold. As a result, we have that there exists some constant c_3 such that

$$\mathbb{P}\left(\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a} \leq c_3 \operatorname{tr}(\boldsymbol{\Sigma}^\star)\right) \geq 1 - \frac{c_2}{2}. \tag{123}$$

By a union bound of (121) and (123), we have

$$\begin{aligned}
&\mathbb{P}\left(\left(\frac{\kappa_y}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \wedge \tau'\right) \left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right| \geq c_1\left(\frac{\kappa_y}{c_3 \operatorname{tr}(\boldsymbol{\Sigma}^\star)} \wedge \tau'\right)\right) \\
&\quad \geq \mathbb{P}\left(\frac{\kappa_y}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \wedge \tau' \geq \frac{\kappa_y}{c_3 \operatorname{tr}(\boldsymbol{\Sigma}^\star)} \wedge \tau'\right) + \mathbb{P}\left(\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right| \geq c_1\right) - 1 \\
&\quad \geq \mathbb{P}\left(\frac{\kappa_y}{\boldsymbol{a}^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}} \geq \frac{\kappa_y}{c_3 \operatorname{tr}(\boldsymbol{\Sigma}^\star)}\right) + \mathbb{P}\left(\left|\langle \boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{U}\rangle\right| \geq c_1\right) - 1 \geq \frac{c_2}{2}
\end{aligned} \tag{124}$$

Redefining constants c_1 and c_2 appropriately, we have

$$\mathbb{P}\left(\left|\langle \widetilde{\boldsymbol{A}}^{\tau'}, \boldsymbol{U}\rangle\right| \geq c_1\left(\frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)} \wedge \tau'\right)\right) \geq c_2, \tag{125}$$

as desired.

D.2. Proof of Lemma 12

We begin by noting that for any matrix $\boldsymbol{U} \in E$,

$$\begin{aligned}\mathbb{E}\left[\sup_{\boldsymbol{U}\in E} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \widetilde{\boldsymbol{A}}_i^{\tau'}, \boldsymbol{U}\rangle\right] &\overset{\text{(i)}}{\leq} \mathbb{E}\left[\sup_{\boldsymbol{U}\in E} \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \widetilde{\boldsymbol{A}}_i^{\tau'}\right\|_{\text{op}} \|\boldsymbol{U}\|_*\right] \\ &\overset{\text{(ii)}}{\leq} 4\sqrt{2r}\,\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \widetilde{\boldsymbol{A}}_i^{\tau'}\right\|_{\text{op}}\right],\end{aligned} \tag{126}$$

where step (i) follows from Hölder's inequality, and step (ii) follows from the definition of the set E. It remains of the proof to bound the expected operator norm in (126). We do this with a trivial modification of the approaches in Vershynin (2010, Section 5.4.1), Tropp (2015, Section 8.6), Kueng et al. (2017, Section 4.1) to accommodate the bounded term $\left(\frac{y+\bar{\eta}_i}{\boldsymbol{a}_i^\top \boldsymbol{\Sigma}^\star \boldsymbol{a}_i} \wedge \tau'\right)$ that appears in each of the matrices $\widetilde{\boldsymbol{A}}_i^{\tau'}$. As a result, there exist universal constants c_1 and c_2 such that if n satisfies $n \geq c_2 d$, then the bound

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \widetilde{\boldsymbol{A}}_i^{\tau'}\right\|_{\text{op}}\right] \leq c_1 \tau' \sqrt{\frac{d}{n}} \tag{127}$$

holds. We conclude by re-defining c_1 appropriately.

E. Proof of Corollary 1

The proof consists of two steps. We first verify that the choices of the averaging parameter m and truncation threshold τ as

$$m = \left\lfloor \left(\frac{(\nu_\eta^2)^2 N}{d}\right)^{1/3} \right\rfloor \quad \text{and} \quad \tau = \frac{y^\uparrow}{\sigma_r r}\sqrt{\frac{N}{md}}, \tag{128}$$

satisfy the assumptions $n \gtrsim rd$ and $\tau \geq \frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)}$. We then invoke Theorem 1.

Verifying the condition on n. We have

$$\begin{aligned} n &= \frac{N}{m} \\ &\overset{\text{(i)}}{=} N\left(\frac{(\nu_\eta^2)^2 N}{d}\right)^{-1/3} \\ &= \left(N^2 \frac{d}{(\nu_\eta^2)^2}\right)^{1/3} \qquad (129) \\ &\overset{\text{(ii)}}{\gtrsim} \left((\nu_\eta^2)^2 r^3 d^2 \frac{d}{(\nu_\eta^2)^2}\right)^{1/3} \qquad (130) \\ &= rd, \qquad (131)\end{aligned}$$

where step (i) is true by plugging in the choice of m from (128), and step (ii) is true by plugging in the assumption $N \gtrsim \nu_\eta^2\, r^{3/2} d$. Thus the condition $n \gtrsim rd$ of Theorem 1 is satisfied.

Verifying the condition on τ. For the term $\sqrt{\frac{N}{dm}}$ in the expression of τ in (128), note that, by the previous point, $\frac{N}{m} = n \gtrsim rd$ (with a constant that, WLOG and by necessity, is greater than 1). Thus $\sqrt{\frac{N}{dm}} \geq \sqrt{r} > 1$. Therefore, it suffices to verify that

$$\frac{y^{\uparrow}}{\sigma_r r} \geq \frac{\kappa_y}{\operatorname{tr}(\boldsymbol{\Sigma}^\star)}. \tag{132}$$

By definition, we have $y^{\uparrow} \geq \kappa_y$. Furthermore, since $\boldsymbol{\Sigma}^\star$ is symmetric positive semidefinite, its eigenvalues are all non-negative and are the same as singular values, and hence $\sigma_r r \leq \operatorname{tr}(\boldsymbol{\Sigma}^\star)$. Therefore, we have (132) holds, verifying the condition on τ.

Invoking Theorem 1. By setting λ_n to its lower bound in (9) and substituting in $n = N/m$ and our choice of τ from (128), we have

$$\lambda_n = C_1 \frac{(y^{\uparrow})^2}{\sigma_r r} \left(\sqrt{\frac{md}{N}} + \frac{\nu_\eta^2}{m} \right). \tag{133}$$

Substituting in our choice of m from (128), we have

$$\lambda_n = C_1 \frac{(y^{\uparrow})^2}{\sigma_r r} \left(\frac{\nu_\eta^2 d}{N} \right)^{1/3}. \tag{134}$$

Substituting this expression for λ_n into the error bound (10) and absorbing C_1 into the constant C, we have

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_F \leq C \left(\frac{\operatorname{tr}(\boldsymbol{\Sigma}^\star)^2}{\sigma_r r} \right) \left(\frac{y^{\uparrow}}{\kappa_y} \right)^2 \sqrt{r} \left(\frac{\nu_\eta^2 d}{N} \right)^{1/3}. \tag{135}$$

Using the fact that $\operatorname{tr}(\boldsymbol{\Sigma}^\star) \leq \sigma_1 r$, we have

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\|_F \leq C \left(\frac{\sigma_1^2}{\sigma_r} \right) \left(\frac{y^{\uparrow}}{\kappa_y} \right)^2 r^{3/2} \left(\frac{\nu_\eta^2 d}{N} \right)^{1/3}, \tag{136}$$

as desired.